nonlinear dynamics

Progenesis LC-MS Tutorial

Including Data File Import, Alignment, Filtering, Progenesis Stats, Protein Search and Protein View

for LC-MS version 2.0

Introduction

This tutorial takes you through a complete analysis of 6 LC-MS runs with 2 groups (3 replicate runs per group) using the unique Progenesis LC-MS workflow. It starts with LC-MS data file loading then Alignment, followed by Analysis that creates a list of interesting features (peptides) which are explored within Progenesis Stats using multivariate statistical methods then onto Protein identity and Reporting.

To allow ease of use the tutorial is designed to start with the restoration of an Archived experiment where the data files have already been loaded. However, the document covers all the stages in the LC-MS workflow, therefore if you are using your own data files refer to Appendix 1 then start at page 5.

How to use this document

You can print this tutorial to help you work hands-on with the software. The complete tutorial takes about 50 minutes and is divided into two sections. This means you can perform the first half focused on LC-MS run alignment and analysis then complete the second half of analysis exploring comparative differences and Protein identity at a convenient time. If you experience any problems or require assistance, please contact us at support@nonlinear.com

How can I analyse my own runs using LC-MS?

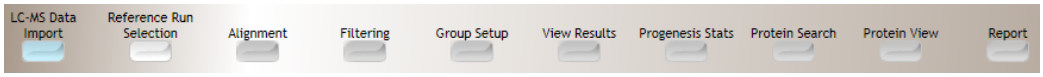
You can freely explore the quality of your LC-MS data using Image QC and then licence your own LC-MS runs using this evaluation copy of Progenesis LC-MS. Instructions on how to do this are included in a section at the end of the tutorial document. Alternatively if you would like to arrange a demonstration in your own laboratory contact support@nonlinear.com and we will help you.

LC-MS Data used in this tutorial

NLD would like to thank Dr Robert Parker and Prof Haroun Shah at the Health Protection Agency, London, UK for providing the example data used in this tutorial as well as invaluable discussion on the handling of the data.

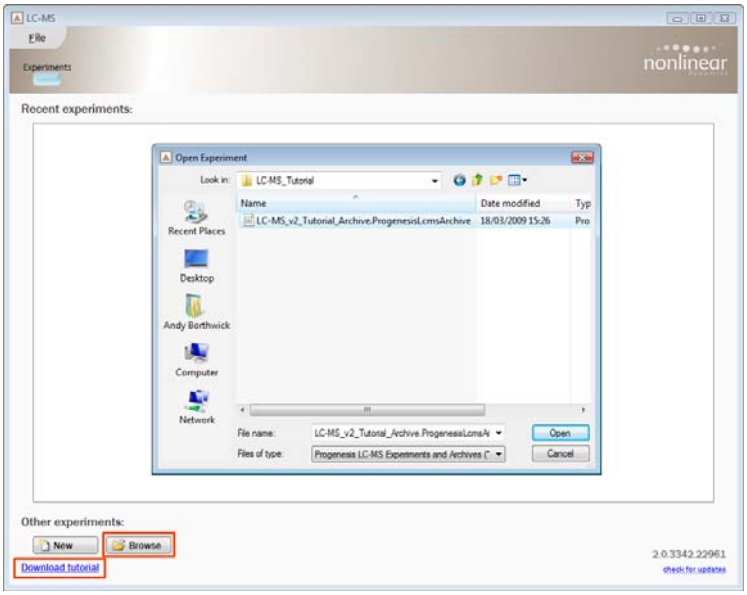
Workflow approach to LC-MS run analysis

Progenesis LC-MS adopts an intuitive **Workflow** approach to performing comparative LC-MS data analysis. The following tutorial describes the various stages of this workflow (see below) focusing mainly on the stages from Alignment to Reporting.



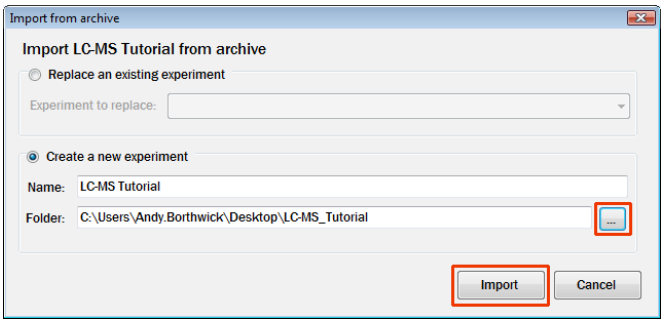
Restoring the LC-MS Tutorial

Open Progenesis LC-MS and downloaded the Tutorial Archive file from the 'Download tutorial' link shown below, placing it in a folder on your desktop . To restore the LC-MS tutorial from this file, locate the LC-MS Tutorial Archive file using the **Browse** button.



This opens the 'Import from Archive' dialog.

Select the **Create a new experiment** option and select the folder in which you placed the archive.

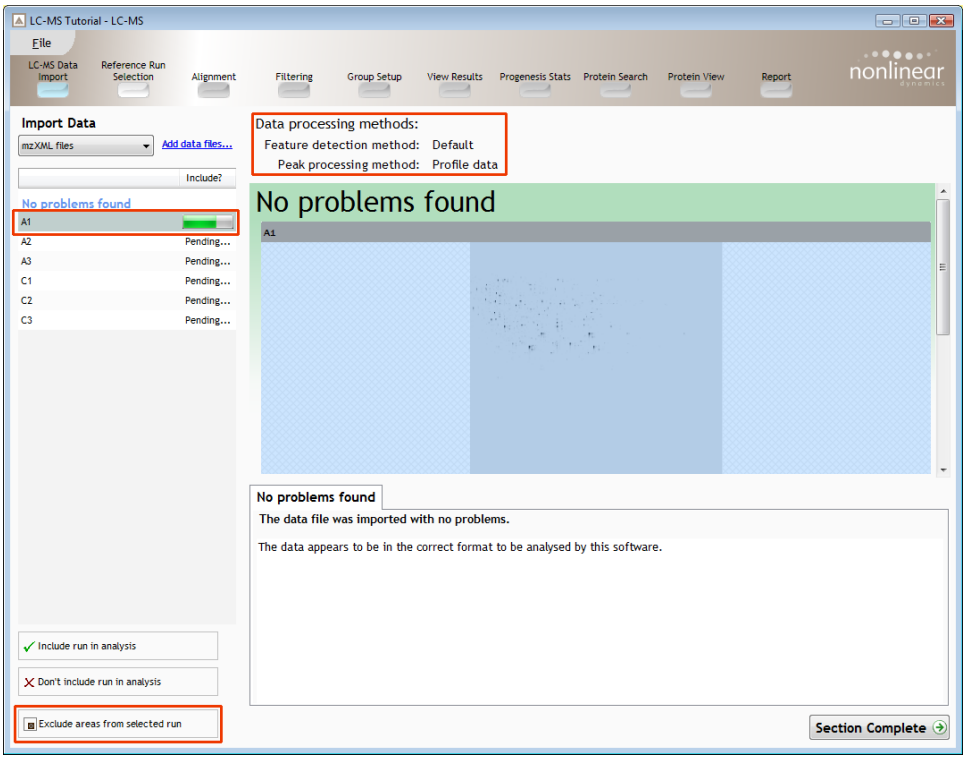


Then press **Import.**

Note: use the **Replace an existing experiment** option if you want to over-write an existing version of the tutorial.

Stage 1: Data import and QC review of LC-MS data set

The LC-MS tutorial will now open at the LC-MS Data Import stage (see below).



Each data file appears as a 2D representation of the run. At this stage you will be warned if any of the data files have been 'centroided' during the data acquisition and conversion process.

Note: as each data file is loaded the progress is reported in the **Import Data** list. The dialog below the image reports on the QC of the imported Data files. In this case 'No problems found' with the this data file.

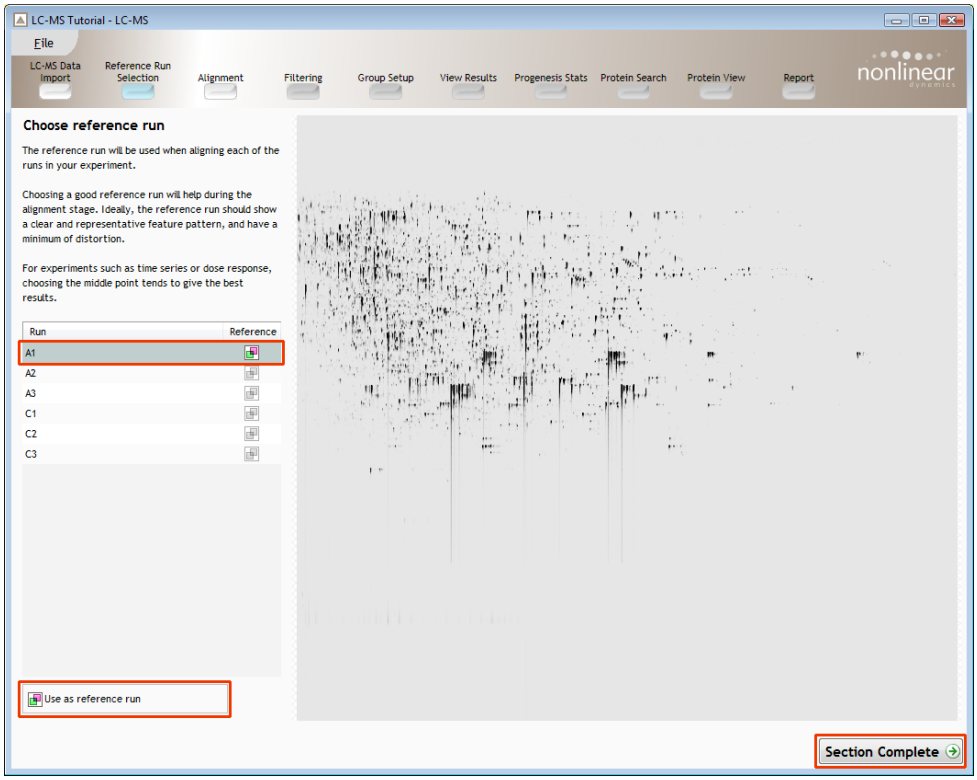
Note: the **'Data Processing Methods'** selected for the Import Data, when the experiment was created, are reported next to the Add data files link (see Appendix 1).

Note: the **'Exclude areas from selected run'** facility allows you to examine and exclude areas (usually early and/or late in the LC dimension (Retention Time) that appear excessively noisy due to capture of data during column regeneration (see Appendix 2). Not required for this data set.

Now move to the next stage in the workflow by clicking **Section Complete.**

Stage 2: Reference data file selection

This stage in the analysis workflow allows you to review and select the most appropriate Reference LC-MS run to align all the other runs to.



To select a Reference run either click on the run in the list and then click **Use as reference run** or double click on the run in the list.

Now move to the next stage in the workflow by clicking **Section Complete.**

Stage 3: Licensing

This stage in the analysis workflow will **always** appear in the LC-MS workflow if you are using 'Unlicensed' data files to evaluate the software and have no dongle attached.

For details on how to use Licensing go to Appendix 3 (page 53)

Stage 4: Alignment

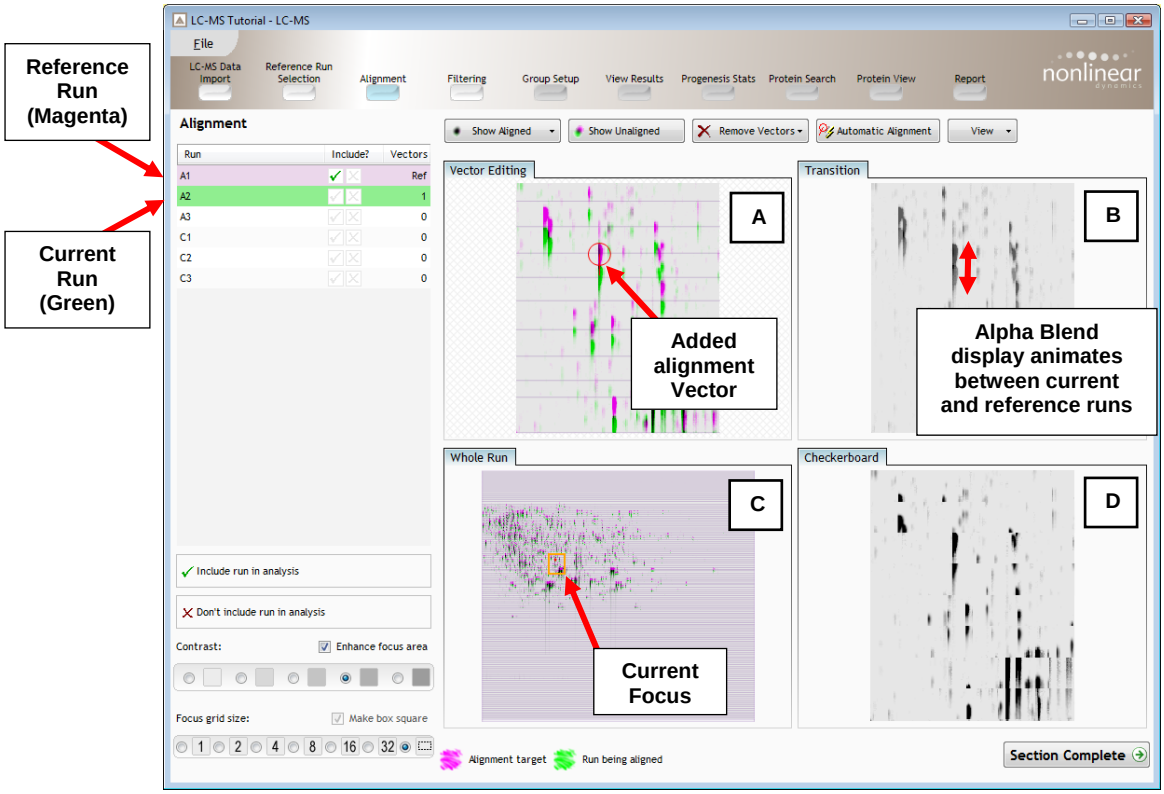
At this stage Progenesis LC-MS Alignment opens displaying your data.

The Program Layout

To familiarize you with Progenesis LC-MS Alignment, this section describes the various graphical features used in the alignment of the LC-MS runs.

To setup the display so that it looks similar to the one below:

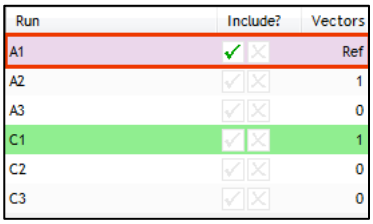
- Click on the features shown in the current focus (orange rectangle) in Window C, this will update windows A,B and D as shown below.
- In window A **click and hold** the left mouse button on a green feature.
- If the green and magenta features (immediately above) have not aligned automatically then **drag** the green feature over the magenta feature and **release** the mouse button.
- The image will 'bounce' back and a red vector, starting in the green feature and finishing in the circled magenta feature will now appear as shown below in window A .



The experiment structure is displayed on the left of the screen in the **Run** panel.

The **Runs:** panel shows the run that is currently being aligned in green, and the run it is being aligned to in magenta.

The **Ref** run for any experiment is the run that you chose to align all the runs to, in this case **A1** highlighted in magenta.



Vector Editing (Window A): is the main alignment area and displays the area defined by the current **focus** rectangle shown in Window C. The current image is displayed in green and the chosen reference image is displayed in magenta. Here is where you place the alignment vectors.

Transition (Window B): uses an **alpha blend** to animate between the current and reference runs. Before the runs are aligned, the features appear to move up and down. Once correctly aligned, they will appear to pulse. During the process of adding vectors, this view will change to show a zoomed view of the area being aligned to help accurate placement.

Whole Data File (Window C): shows the **focus** for the other windows. When you click on the view the orange rectangle will move to the selected area. The focus can be moved systematically across the view using the left and right cursor keys. The focus area size can be altered using the controls in the bottom left of the screen or by clicking and dragging out a new area with the mouse.

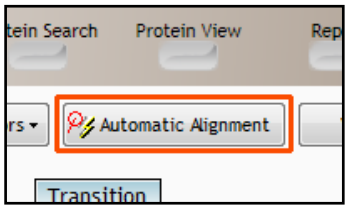
Checkerboard (Window D): shows a checkerboard view of the current run interleaved with the reference run. When the area is aligned the edges of the squares start to merge.

The unique use of the alpha blend and checkerboard views helps to make highly accurate alignment of feature borders more obvious to the human eye.

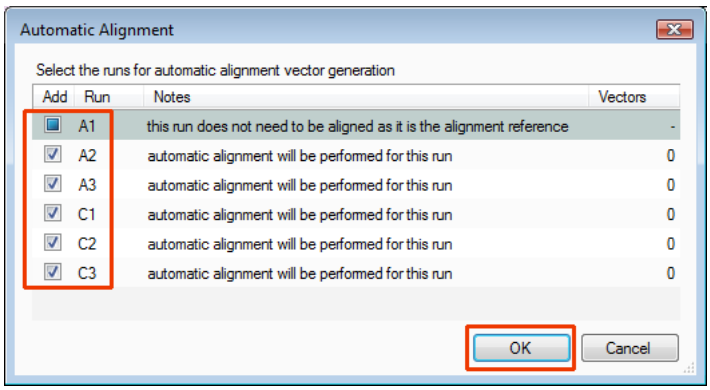
Generation of alignment vectors

The alignment of LC-MS runs is required in the LC (retention time) direction, this is key to correcting for the variable elution of peptides during the chromatographic separation.

The Alignment algorithm will generate 'Automatic' vectors, in the retention time direction for each run, to enable the alignment of all the LC-MS runs to the 'Reference Run'.



The alignment vectors are generated automatically for all the LC-MS runs by using the 'Automatic vector wizard' accessed by clicking on **Automatic Vectors** on the top tool bar.



Select (tick) the runs you require to generate vectors for and click **OK**.

If applying alignment automatically now move to page 15

In some cases, where the misalignment is severe, using a combination of a 'few' manually placed vectors on each run and then using the Automatic vector wizard to generate the rest of the vectors for each run can give better results.

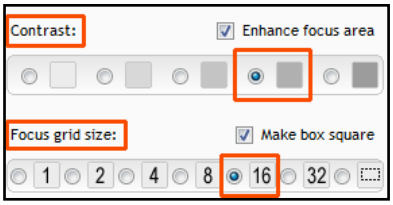
In this example try placing some manual vectors before generating the automatic vectors.

The following section describes how to manage the placement of manual vectors on your LC-MS runs

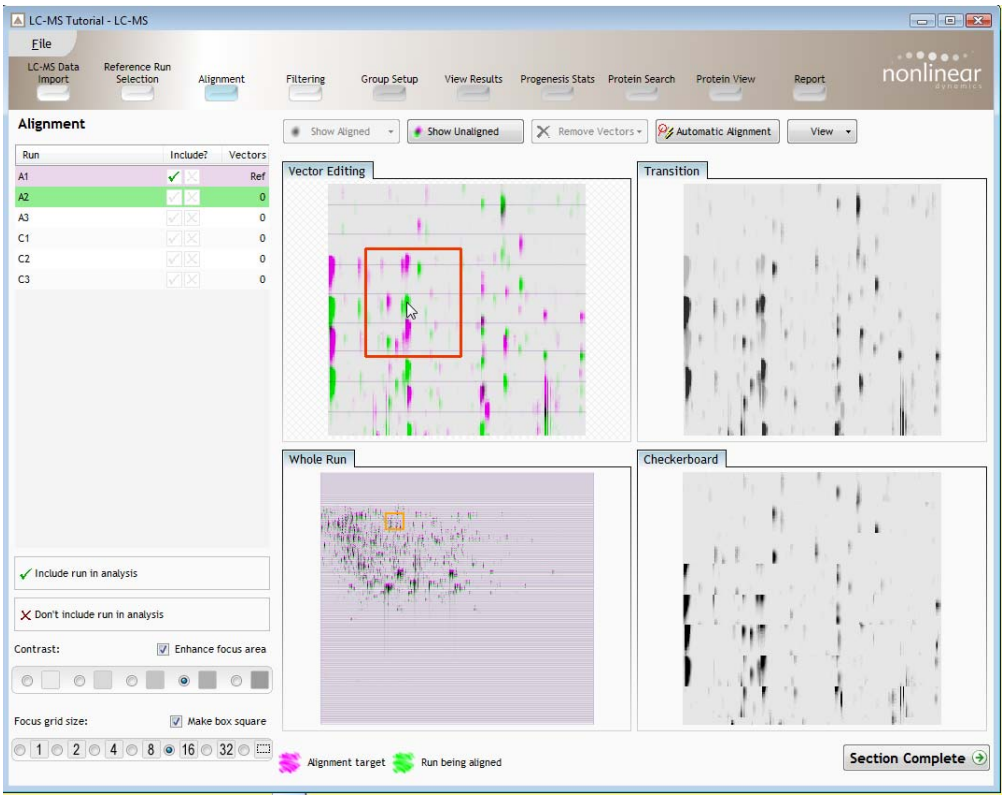
Approach to alignment

To place manual alignment vectors on a run (A2 in this example):

1. Click on Run A2 in the **Runs** panel, this will be highlighted in green and the reference run (A1) will be highlighted in magenta.
2. You will need approximately 5 - 10 **alignment vectors** evenly distributed from top to bottom of the whole run.
3. First ensure that the size of the focus area is set to **8 or 16** in the Focus grid size on the bottom left of the screen.

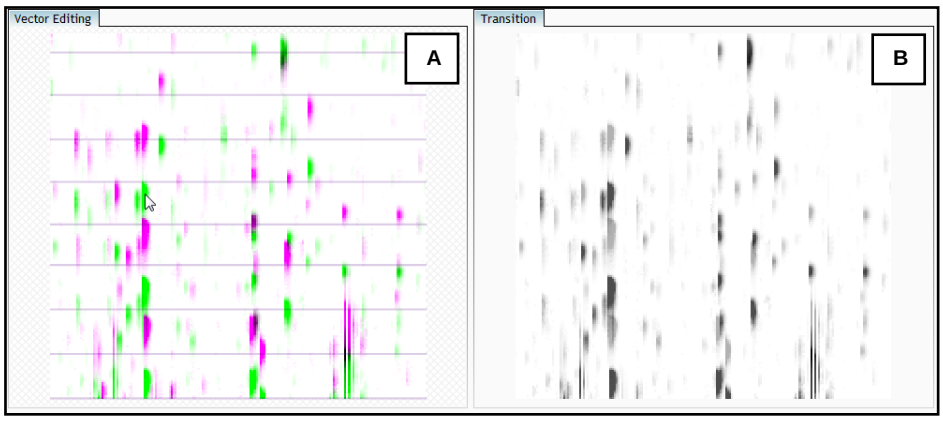


4. Click on an area (see below) in the **Whole Run** window (C) to refocus all the windows. Adjust Contrast as required.

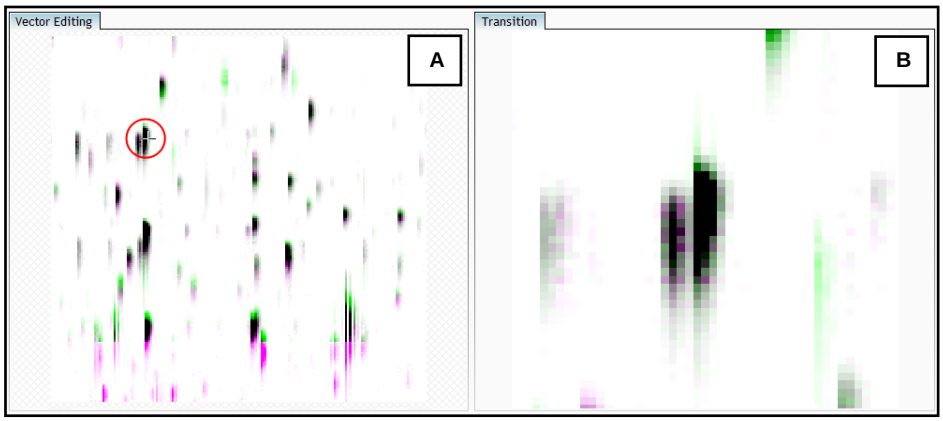


Note: the features moving back and forwards between the 2 runs in the **Transition** view indicating the misalignment of the two LC-MS runs

5. Click and hold on a green feature in Window A as shown below.

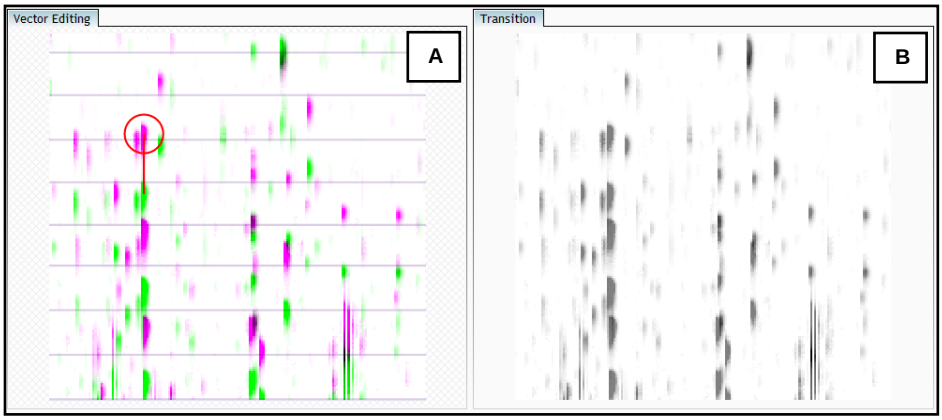


6. As you are holding down the left mouse button drag the green feature over the corresponding magenta feature of the reference image. The vector will appear as shown below as a red circle with a 'cross hair' indicating that a positional lock has been found for the overlapping features.



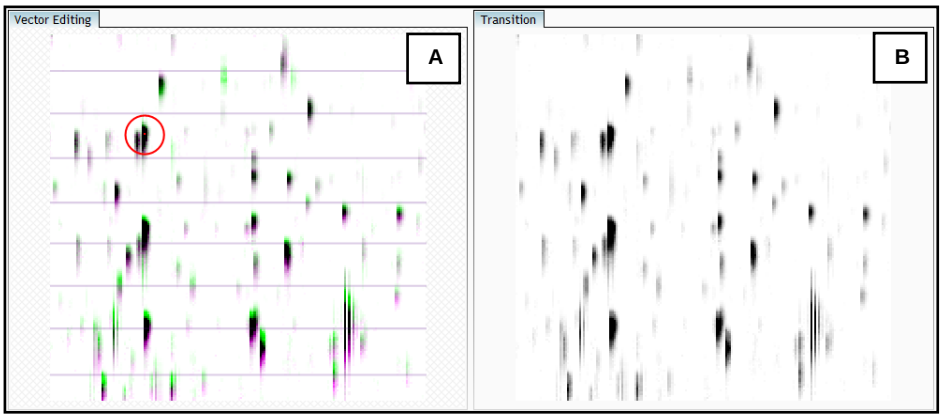
Note: as you hold down the mouse button, window B zooms in to help with the alignment.

7. On releasing the left mouse button the image will 'bounce' back and a red vector, starting in the green feature and finishing in the magenta feature will appear.

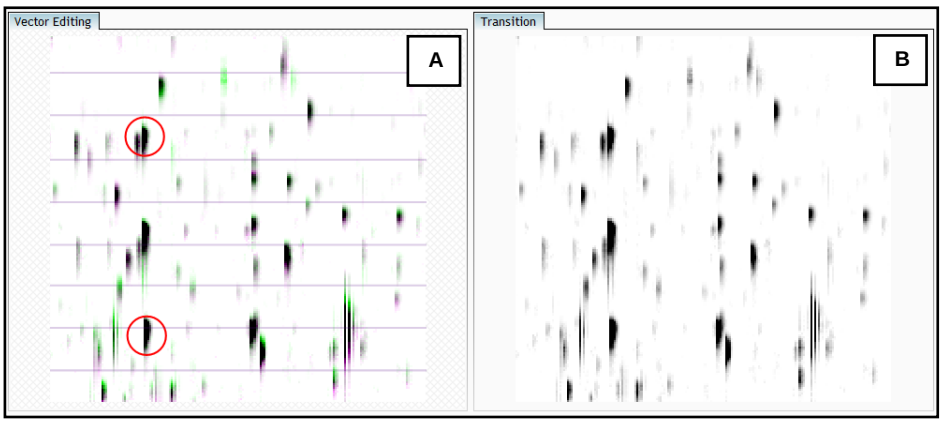


Note: an incorrectly placed vector is removed by right clicking on it in the **Vector Editing** window

8. Now click Show Aligned on the top tool bar to see the effect of adding a single vector.

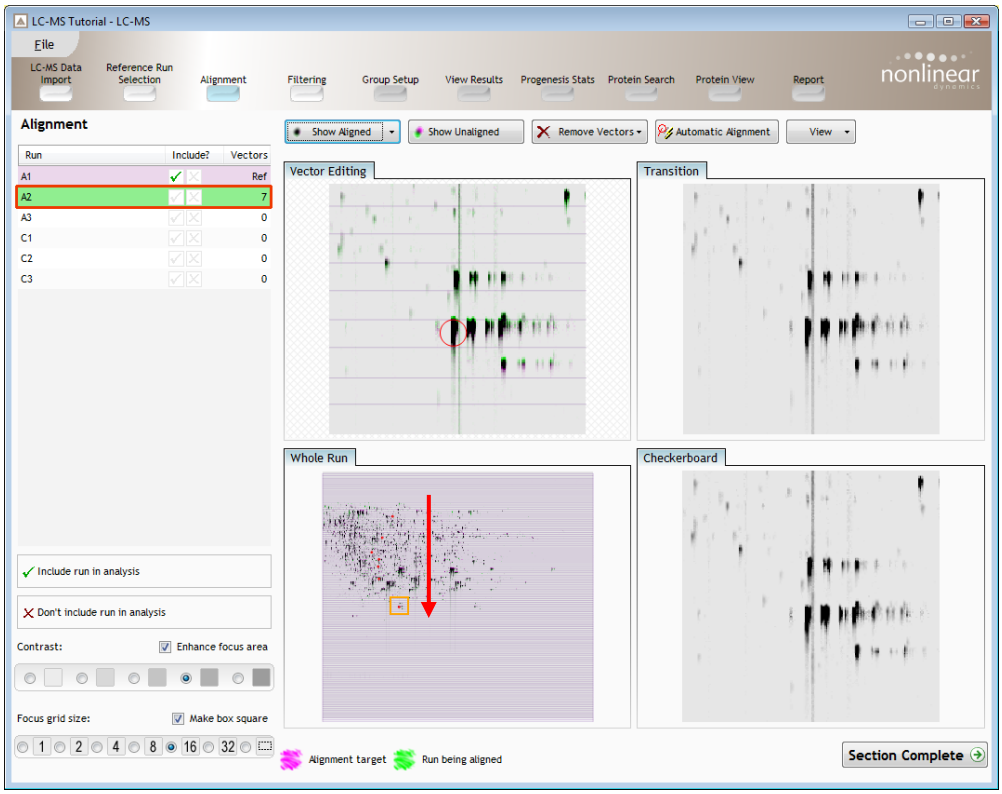


9. Additing an additional vector will improve the alignment further. **Note** this time as you click to add the vector it 'jumps' automatically to the correct position using the information from the existing alignment vector



10. Repeat this process moving the focus from top to bottom on the **Whole Run** view

Note: the number of vectors you add is recorded in the **Runs** table

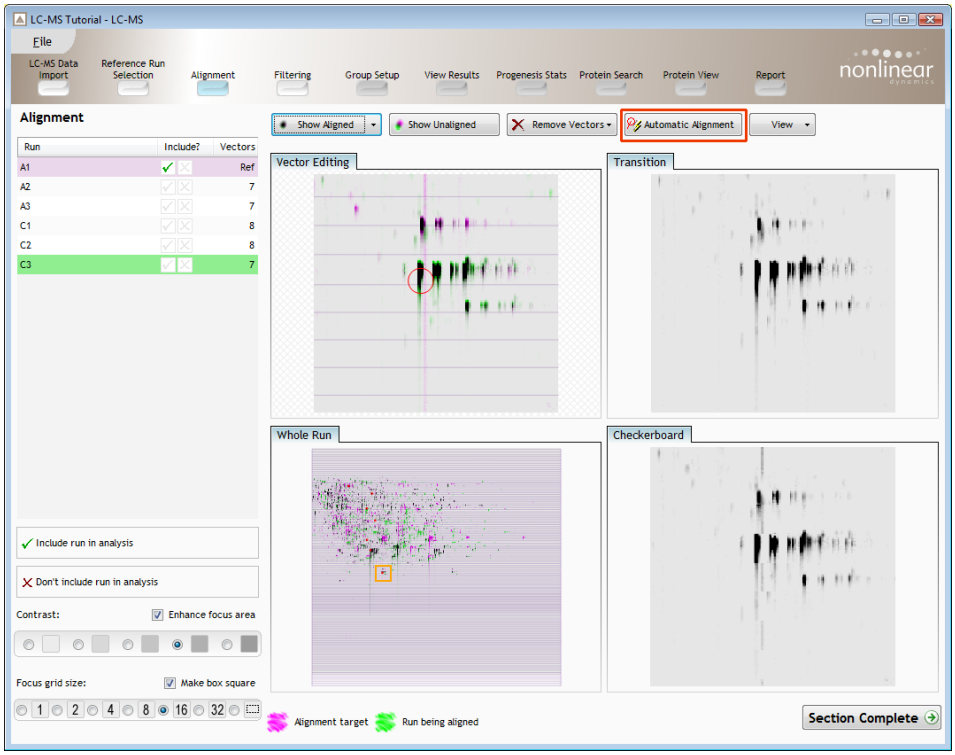


11. Now move on to the next run to align and repeat the addition of a few manual vectors

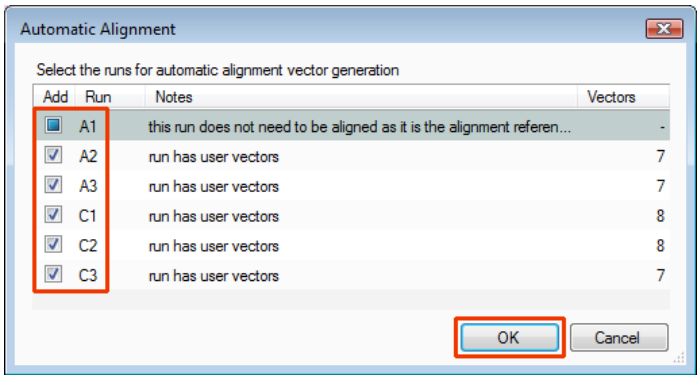
The number of manual vectors that you add at this stage is dependant on the misalignment between the current run and the Reference run. In many cases only using the Automatic vector wizard will achieve the alignment.

Also the ‘ease’ of addition of vectors is dependant on the actual differences between the LC-MS runs being aligned

12. Repeat this process for all the runs to be aligned.



13. Then select Automatic vectors and click **OK**.

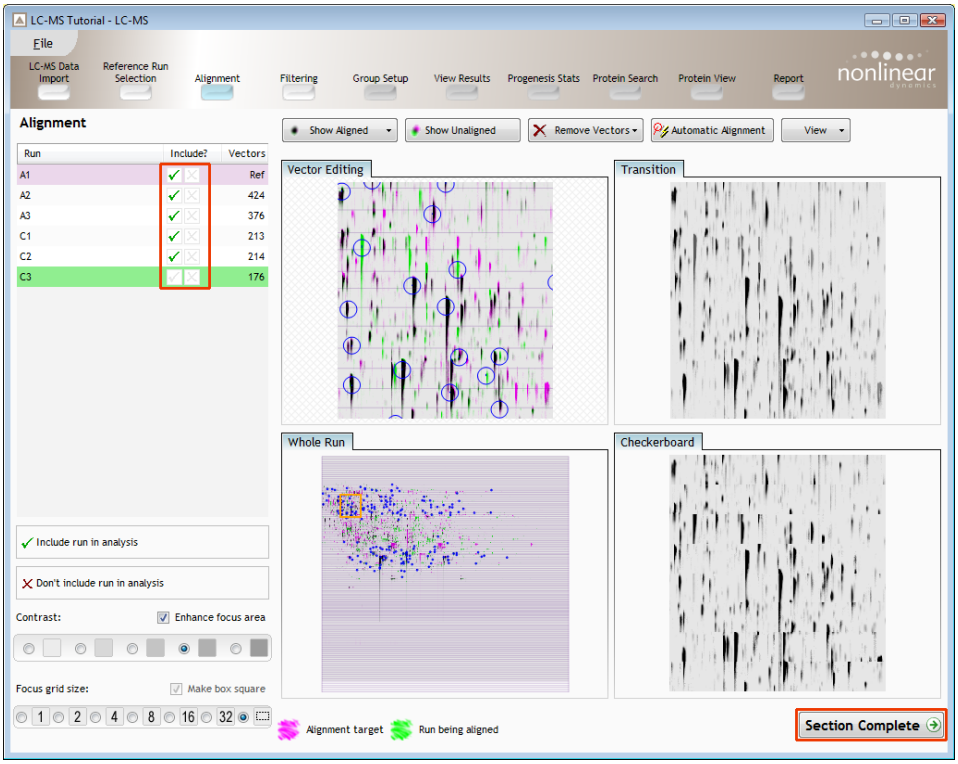


Note: the tick boxes next to the 'Run name control' which control whether vectors will be generated for each run.

Automatic generation of alignment vectors

After applying **Automatic alignment** the number of vectors will be updated on the **Runs** panel and the vectors will appear (in blue) on the image.

If the alignment has worked well then in Windows A and C the grid lines should show minimal distortion, Window B will show features pulsing slightly but not moving up and down.



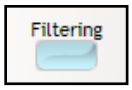
At this point, you should check the automatically placed (blue) vectors. This will be easier with a larger grid size. Make sure the grid size is set to 4 using the **'Focus grid size'** control at the bottom left of the window.

In each square, you can, if required edit the vectors to improve the image alignment.

To indicate that alignment of the run is complete and ready for Analysis click **Include Run in analysis** located underneath the image list.

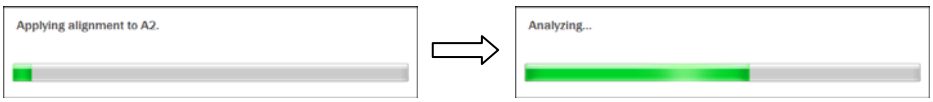
For this tutorial example we are **NOT** performing any editing of the Automatic alignment.

Stage 5: Filtering

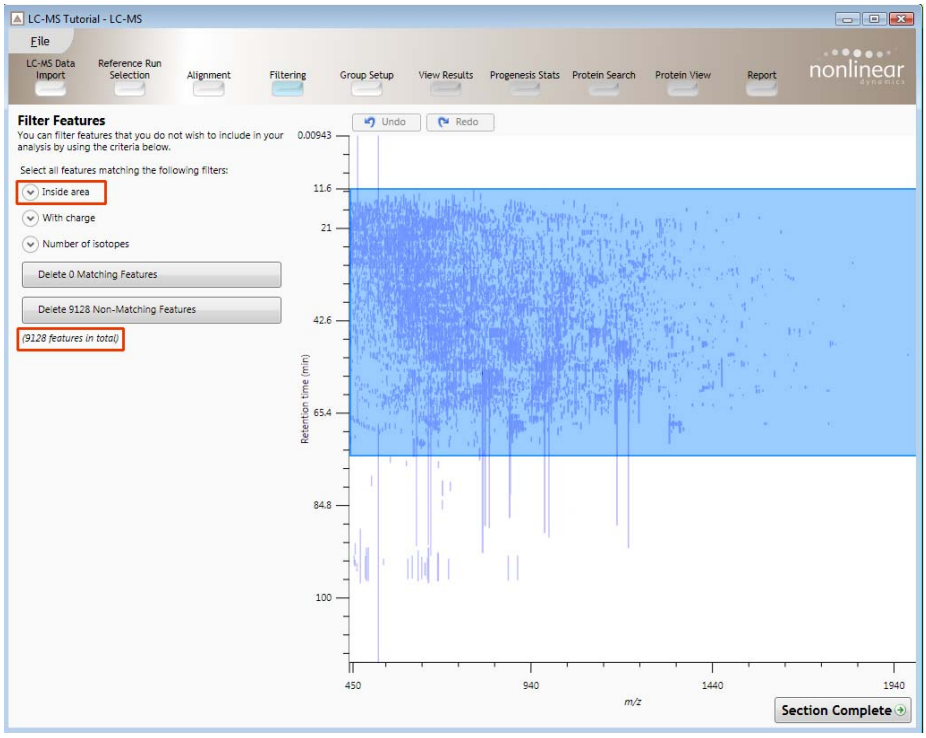


Now that you have reviewed your aligned Runs, you are ready to analyse them. Move to the **Filtering** stage, by either clicking on **Section Complete** (bottom right) or on Filtering on the workflow.

During the few minutes that the automatic analysis requires, a progress bar will appear telling you first that it is applying alignment to the Runs and then that it is Analysing.

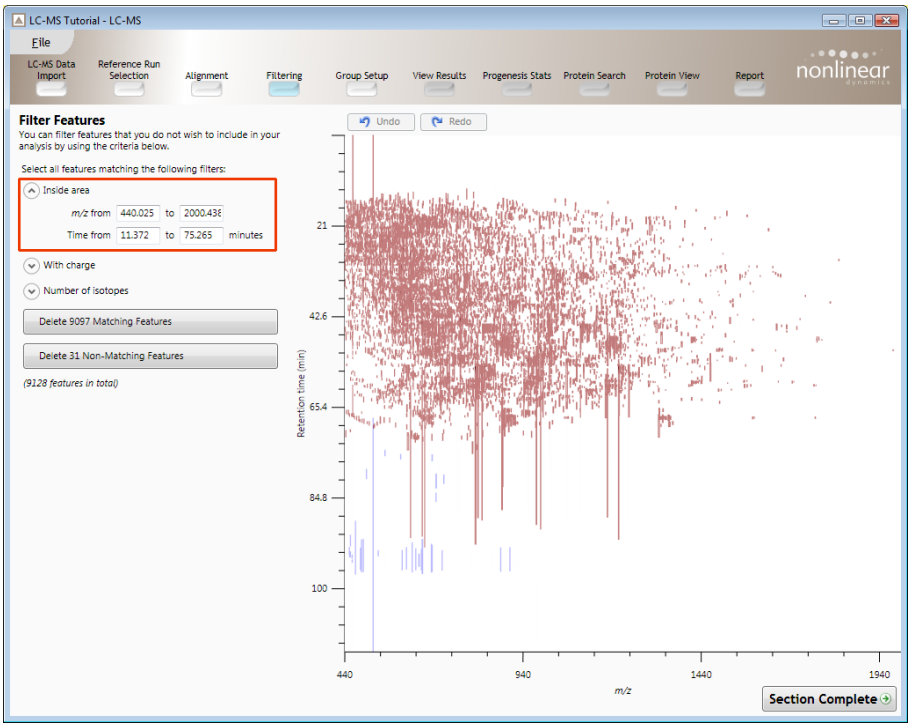


On completion of analysis the Filtering stage will open displaying the number of features detected in this example 9128. If required you can remove features based on position, charge state, number of isotopes or combinations of these feature properties.



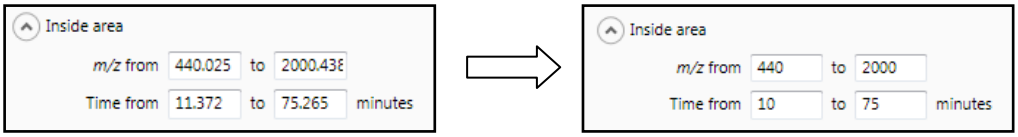
For example, to delete features with early and late ‘Retention times’ drag out an area as shown.

All features with their reported ‘Retention Time and m/z’ values contained within the mask will be selected.



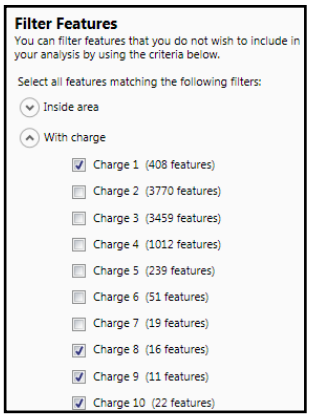
As you release the mouse button the values for the masked area will appear on the top left

Note: the limits can be adjusted by entering the required values in the boxes



To remove the (in this case 31) features outside of the selected area, press the **Delete 31 Non-Matching Features** button

In addition to setting limits for 'Retention time and m/z', features can also be selected on the basis of charge or the number of isotopes present. Thus allowing you to refine the selection through a combination of feature properties



For example: when charge state is selected the number of features present at each charge state is displayed, these can be selected accordingly.

Area limits, charge state and number of isotopes can be combined to refine the feature selection.

For this tutorial, we will filter the area as shown and remove features with a charge state of 1 and 8 and above.

We will now delete a further 547 with a charge state of 1 and 8 and above by ticking the various options.

Hence all features with a charge state of 1 and 8 and above will appear red (see above).

To remove these features with press **Delete 547 Matching Features**

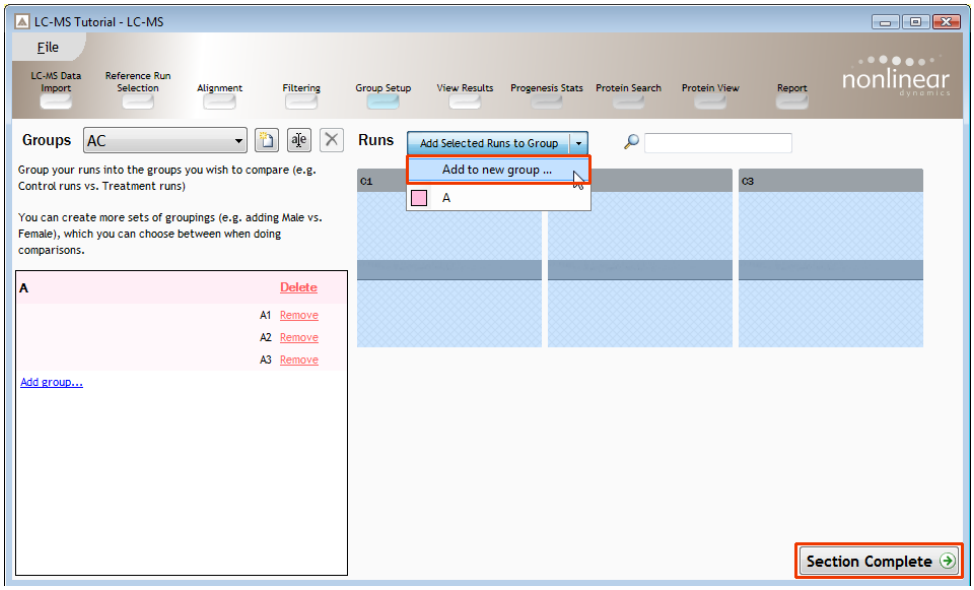
You can use the **Undo** button to bring back deleted features, however, when you move to the next section you will lose the capacity to undo the filter.

To move to the next stage in the workflow click **Section Complete.**

Stage 6: Group Setup for Analysed LC-MS runs

At this stage in the workflow you can setup one or more groupings of your sample data.

For this example, group the analysed LC-MS runs to reflect the Biological groupings in the original study. This tutorial contains 2 groups: A and C, with 3 replicates each.



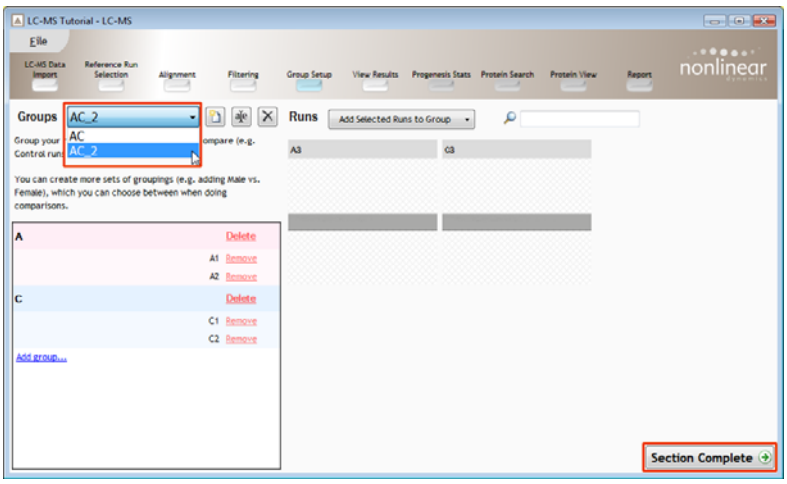
Creating a group

1. Select the Runs in a group (based on names at top of image) by clicking on the required thumbnails (Highlighted as above)
2. Press the 'black triangle' next to the **Add Selected Runs to Group** button on the main toolbar.
3. Select **Add to new group…** from the **drop down menu**.
4. A new group will appear in the **Groups** pane on the left panel
5. Rename the group by over typing the new group name (e.g. A)
6. Repeat steps 1 to 5 until all the Runs are grouped.

In the example shown the grouping has been renamed "AC" using the rename button

To create another Group Setup, for example comparing only 2 replicates for A and C groups, click on Create a new group setup (see right).

Give it a new name (i.e. AC_2). The Runs will reappear in the main window. Create the new groups as described above.



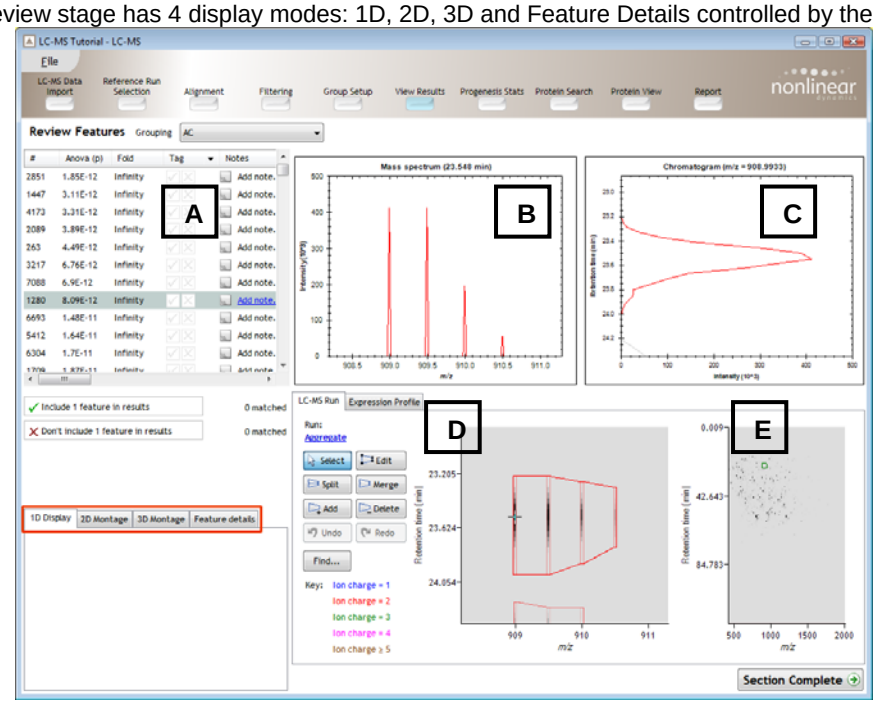
Note: the **Group set up** drop down will now contain both setups and the ungrouped data files (A3 and C3) will remain in the main window.

To move to the next stage in the workflow click **Section Complete.**

Stage 7: Validation, review and editing of results

The purpose of this stage in the Workflow is to review the list of features using the visual tools provided and edit features if required.

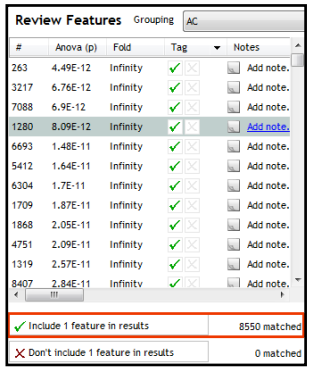
The review stage has 4 display modes: 1D, 2D, 3D and Feature Details controlled by the tabs on



the bottom left of the display. Each display has multiple views to allow comparative exploration of the detected features on the aligned LC-MS runs.

Exploring analysed data using the Data displays

The 1D Display

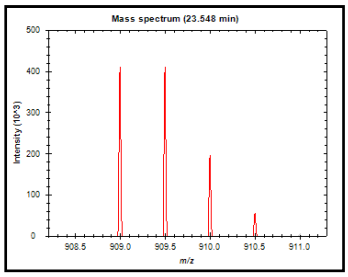


Window A: shows the list of features ranked by the p value for the one way **Anova** using the current grouping.

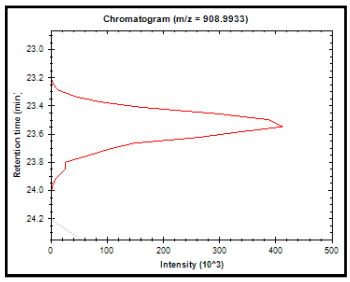
Note: A value of 'Infinity' in the **Fold** column indicates 'Presence/Absence'

To include a feature in the selection for the next section of the analysis, click on the **Include features in results** button at the bottom of the table. On clicking the button it will move on to the next feature on the list.

To select a group of features drag out a selection on the table and click on the **Include feature in results** button (see right)

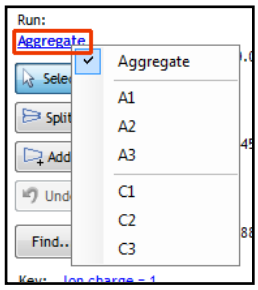


Window B: displays the Mass spectrum for the current feature on the selected Run (in window D).



Window C: displays the Chromatogram for the current feature on the selected Run (in window D).

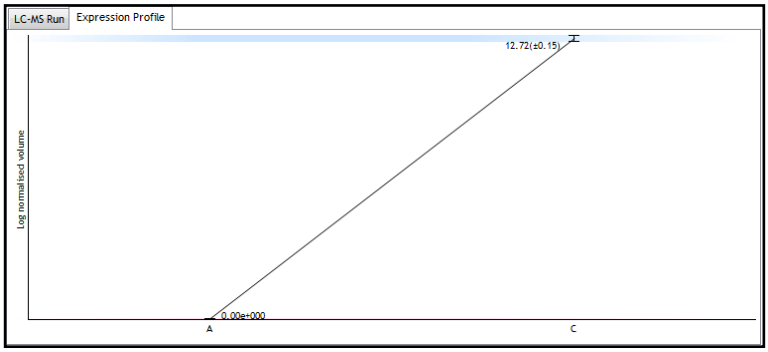
Window D: displays the details of the LC-MS currently selected run. By default the information is displayed as an Aggregate for the feature across all the aligned runs.



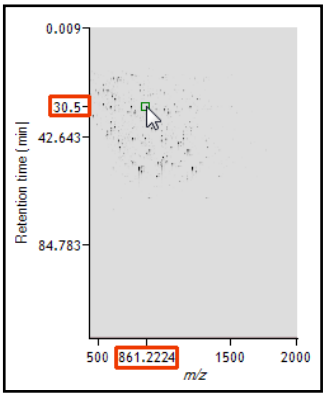
Details of individual runs can be viewed by using the 'Run' link and selecting the run you wish to view.

The feature editing tools are located in this window (see page 25 for functional explanation).

Clicking on the Expression Profile tab in Window D shows the comparative behaviour of the feature across the various biological groups based on group average normalised volume. The error bars show +/- 3 standard deviations.



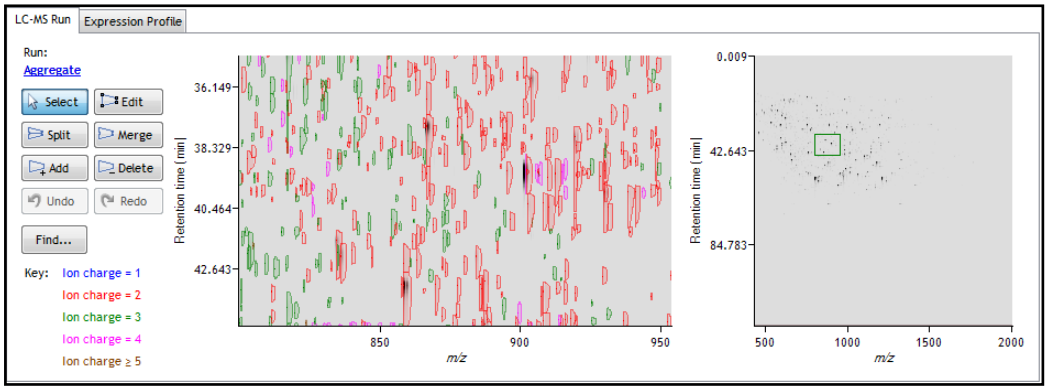
Window E: shows where the current feature is located on the LC-MS run by means of the 'Green' rectangle.



To change the current location, click on the image of the run (note: the retention time and m/z values update as you drag the cursor around this view).

Note: doing this updates the focus of all the other windows.

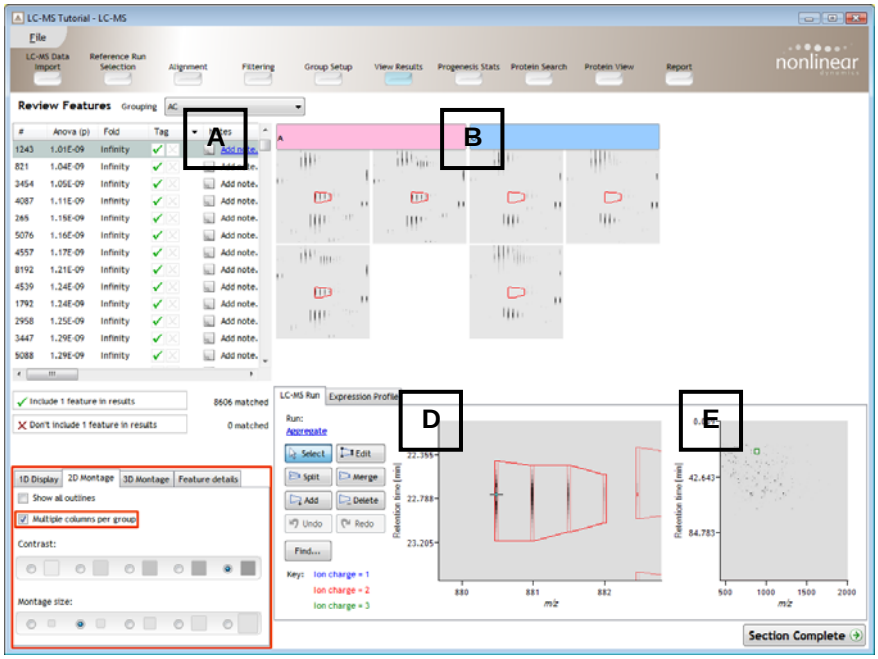
You can also drag out a larger area on this view that will refocus the other windows



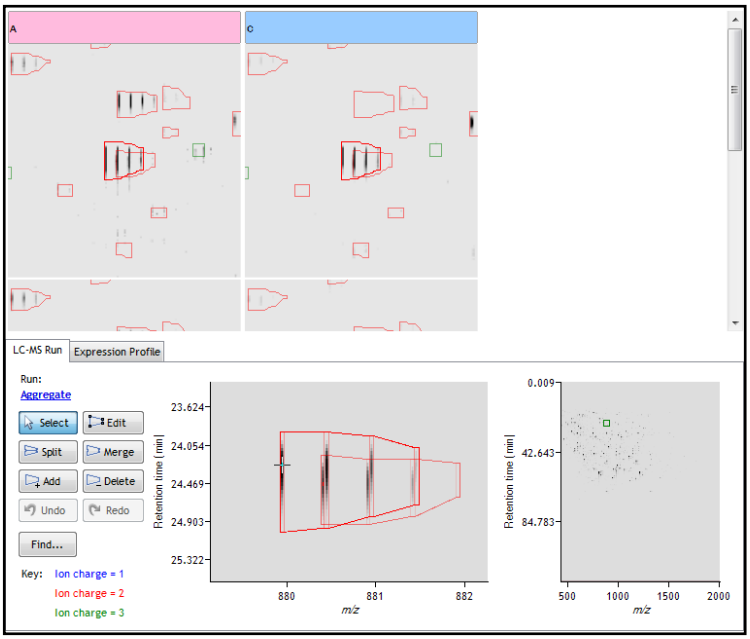
The 2D Display

Windows A, D and E: perform the same functions across all 4 display modes.

In the 2D Montage mode, Window B displays a montage of the current feature across all the aligned LC-MS



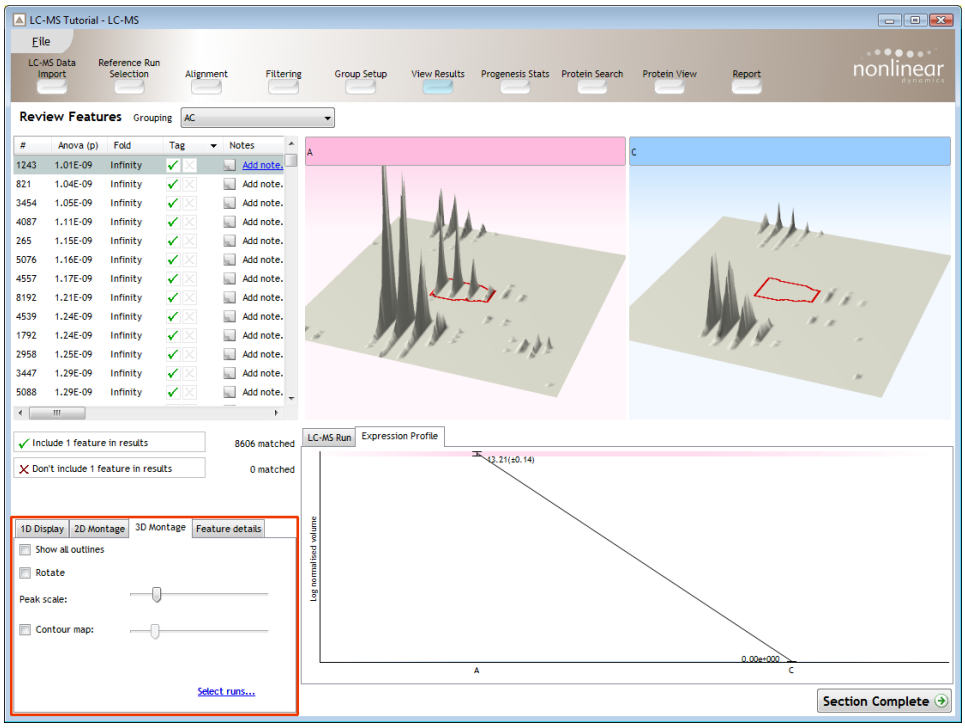
The appearance of the Montage (window B) is controlled by the panel on the bottom left of the display.



Using the the various views in the 2D display one can examine the feature detection in detail to validate the correct detection of even fully overlapping features as shown above.

The 3D Display

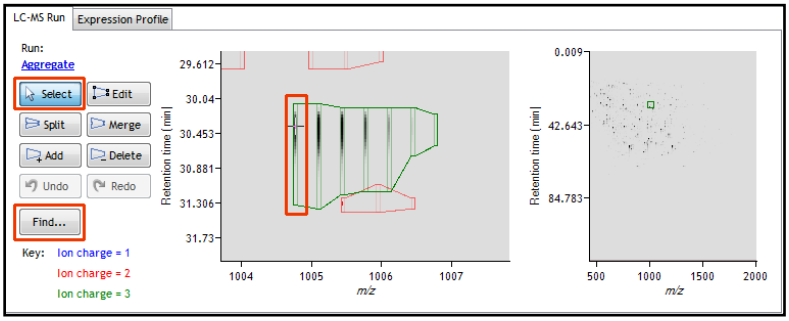
Window B changes into a 3D view by selecting the 3D Montage tab on the bottom left of the display.



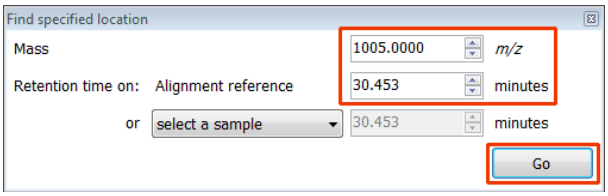
The number of 3D views displayed in the montage is controlled using the Select runs link on the 3D Montage tab. The images can be set to **Rotate** automatically or you can rotate them manually by clicking and dragging them with the mouse.

Editing of features in the View Results stage

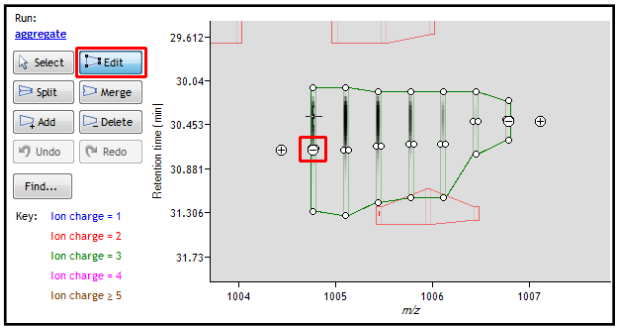
As an example of using the editing tools which are located on the left of the LC-MS Run view, we will remove and add back the 'monoisotopic peak' for the detected feature selected below. A feature can be selected from the 'Features' list or located using the various image views .



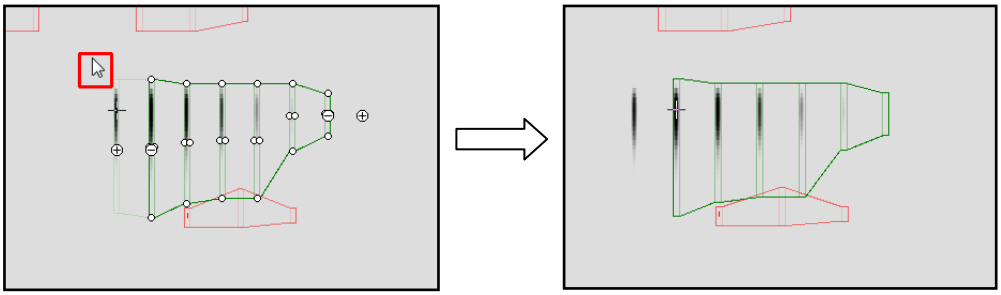
1. Locate the feature at approx 1005 m/z and 30.453 min using the **Find** tool.



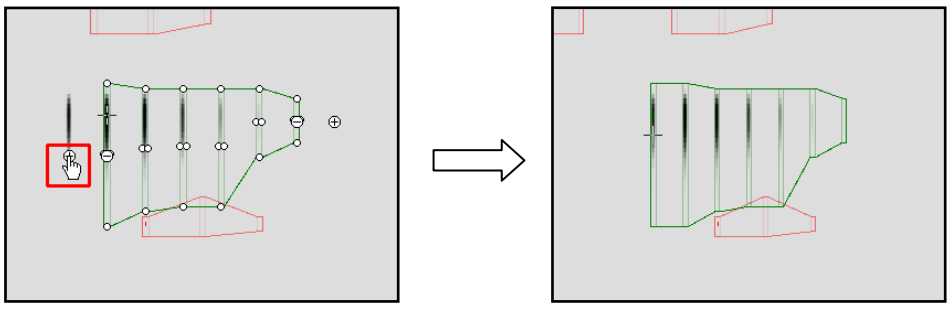
2. Select the **Edit** tool and click on the feature to reveal the 'edit handles'



3. Click on the 'minus' handle over the monoisotopic peak to remove it.
4. Click outside the boundary of the feature to update the view.



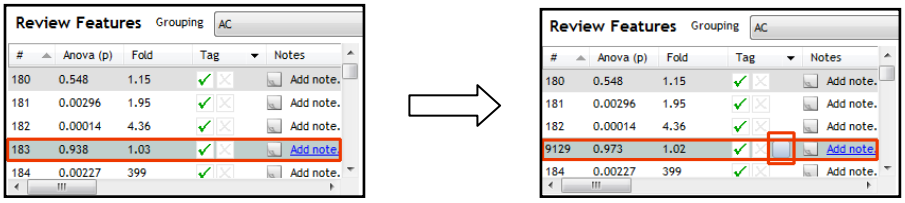
5. To add a peak to an existing feature, ensure that **Edit** is selected then click inside the feature to reveal the handles.



6. Click on the 'plus' handle on the peak to add it.

7. Then click outside the feature to update the view.

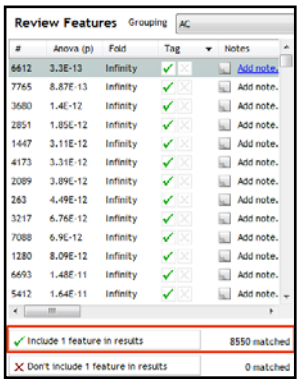
8. **Note**: If you are not satisfied with the editing use the **Undo** button and retry.



9. Finally note: that a tag is automatically added to the edited feature in the table and the features id number is changed to the next available one in the list.

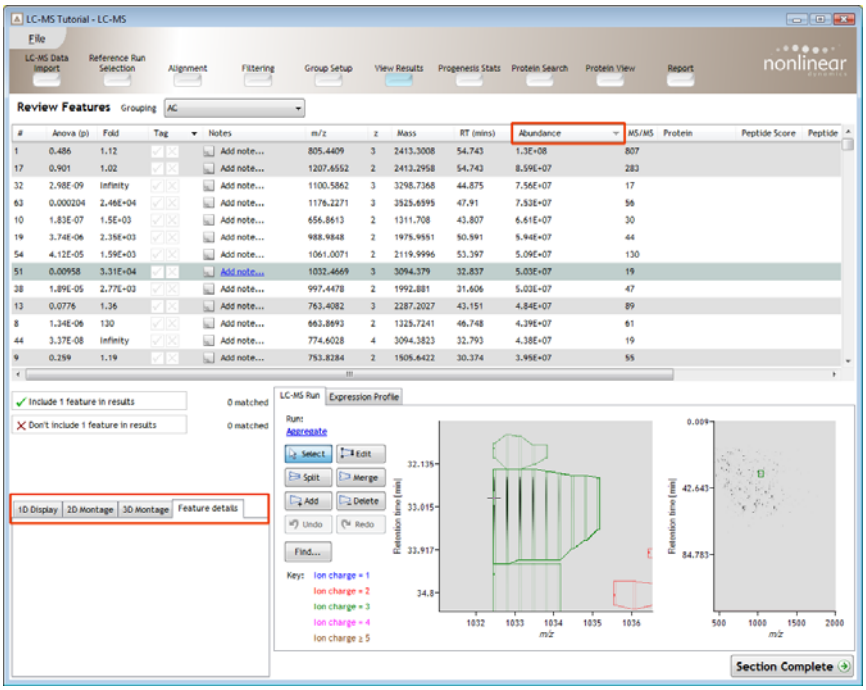
The other tools: split, merge, add and delete behave in a similar fashion and their use can be combined to achieve the desired results.

Selecting and tagging features for Progenesis Stats

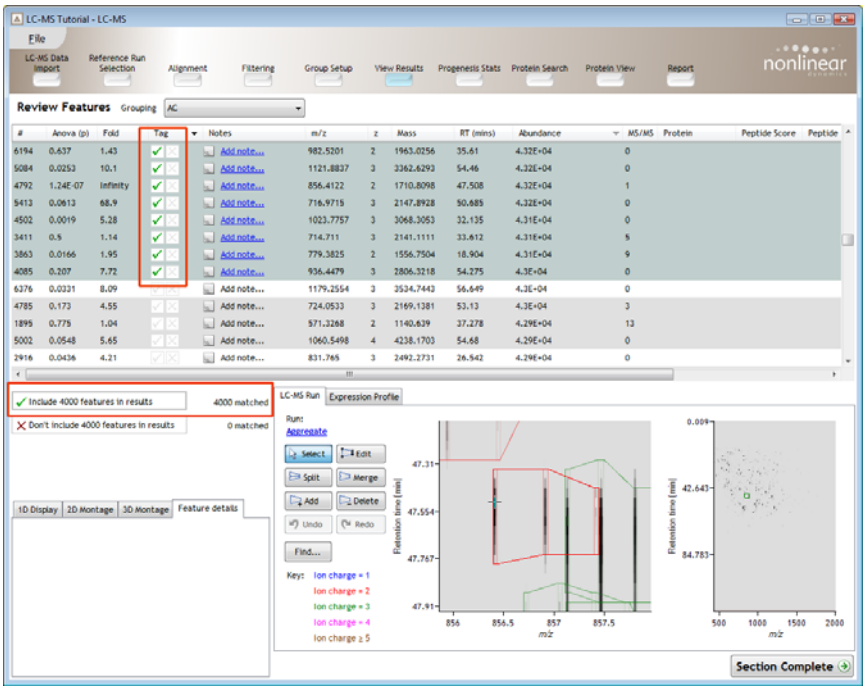


There are a number of ways to 'refine' your 'Ranked List' of analysed features before examining them with the Statistical tools in Progenesis Stats. These make use of simple 'Selection' and 'Tagging' tools that can be applied to the various Groupings created in Stage 6 (page 19). An example is described below.

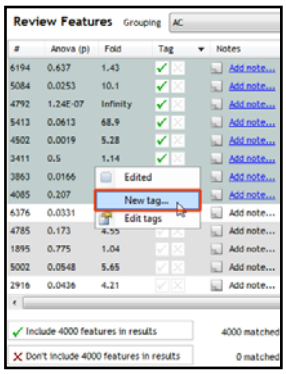
First expand the 'Features' table to show all the details by clicking on the 'Feature details' tab on the bottom left of View Results. Then order on Abundance and select the top 4000 features



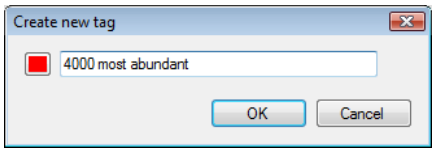
Now click on the 'Include 4000 features in results' this will place a tick against each highlighted feature.



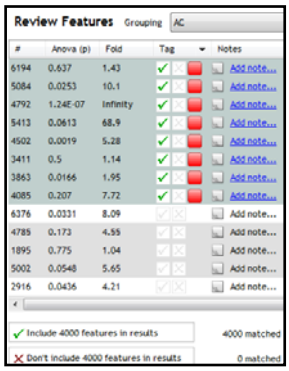
With the 4000 features still highlighted right click on them and select 'New Tag'



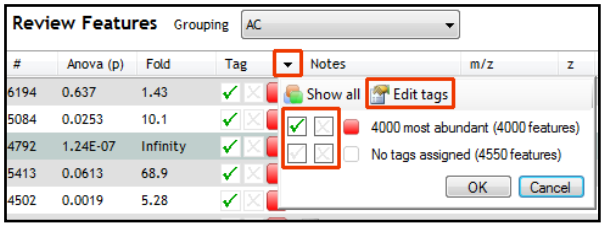
Give the Tag a name. i.e. **'4000 most abundant'**.



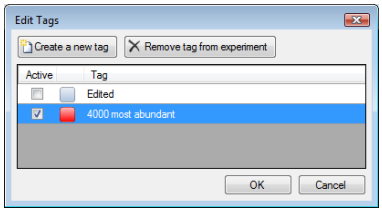
On clicking **OK** the Tag is applied to the features highlighted in the table (signified by a coloured square).



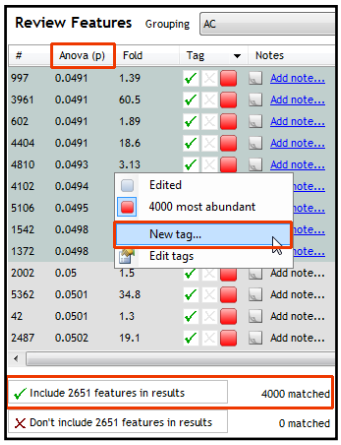
To view the Tags and also control the number of features displayed in the table, click on the drop down selection on the right of the Tag column header. By ticking one or more Tags you can control the number of features displayed.



To delete and/or create additional Tags click on **Edit tags** and Create/Remove Tags as required.

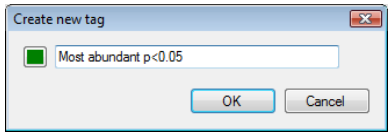


Now with the table displaying the 4000 tagged features order them by Anova (p value) and highlight all the features with a value less than 0.05

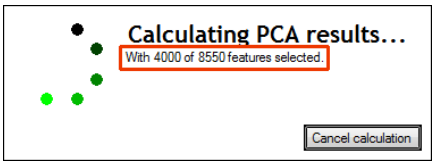


Right click on the features and create a 'New Tag' called **Most abundant p<0.05.**

Press OK to add this new tag to 2651 features



To move to the next stage in the workflow click **Section Complete.**



Note: as Progenesis Stats opens progress dialog will appear as the Stats are calculated for the features selected at the View Results Stage, in this example 4000 features

Stage 8: Multivariate Statistics on Selected Features

The tutorial displays the functionality of the Multivariate Statistics. This section is only available if Progenesis Stats is licensed .For this tutorial we start by examining the behaviour of the 4000 selected features from the previous stage, **View Results**, then making use of the Tags and Group Setups to explore specific groups of features that meet chosen statistical thresholds.

The statistical analysis of the selected data is presented to you in the form of interactive graphical representations of answers to questions asked of the analysed data.

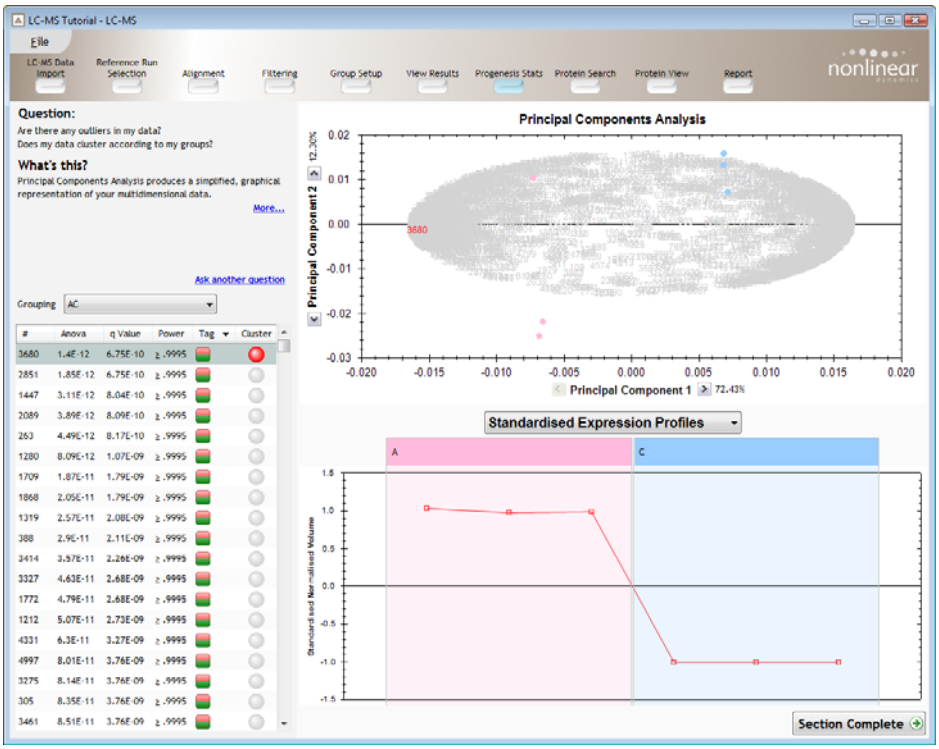
Principal Component Analysis (PCA)

On opening **Progenesis Stats** the statistically based question asked of the application is:

Are there any outliers in my data? And does my data cluster according to my groups?

It answers this question by:

'Using Principal Components Analysis (PCA) to produce a simplified graphical representation of your multidimensional data'.



Note: the LC-MS runs (samples) are displayed as solid coloured circles on the plot. To identify the runs, a tooltip is displayed when the cursor is held over each circle.

PCA can be used to determine whether there are any outliers in the data and also look at how well the samples group. The groupings that can be observed on the 2D PCA plot can be compared to your experimental design groupings and conclusions can be drawn regarding possible outliers in your original groupings.

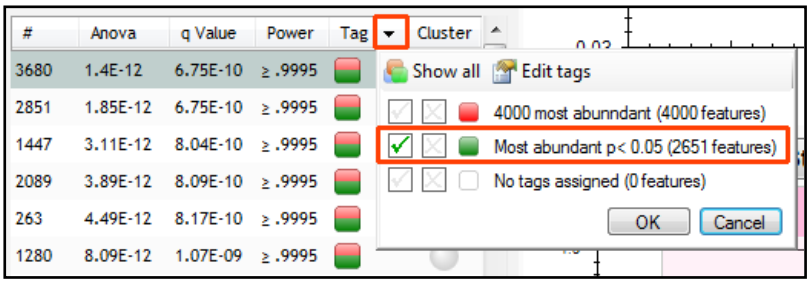
Correlation Analysis

The Table in the Stats view contains additional columns:

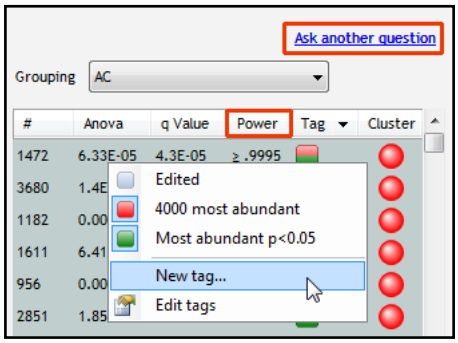
q value: tells us the expected proportion of false positives if that feature's p-value is chosen as the significance threshold

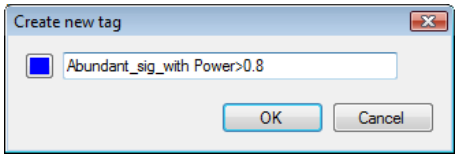
Power: can be defined as the probability of finding a real difference if it exists. 80% or 0.8 is considered an acceptable value for power. The Power Analysis is performed independently for each feature. Given the sample size, the variance of expression values and the difference we want to detect, we can calculate the power. Also, for a given power of 80% we can determine how many samples are required to ensure we find a difference if it actually exists.

Use the tags created in View Results filter the features in the table. In this example using the features tagged **'Most abundant p<0.05'** (2651).

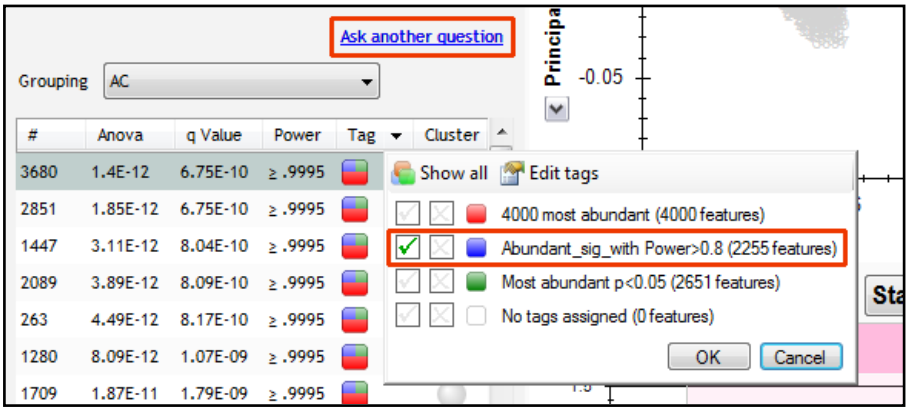


Now set a tag for all features with a power of 0.8 or greater by first ordering the table on Power and then highlighting all the features with a value of 0.8 or greater, right click on the highlighted features and select '**New tag**'





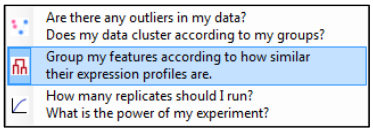
Having created the new tag, now use it to filter your data to display only those features with a Power value greater than 0.8



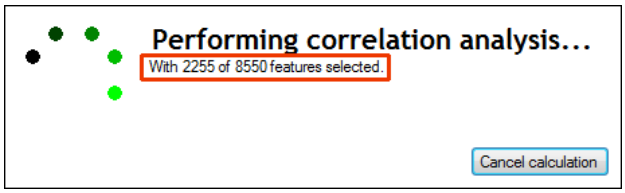
Correlation Analysis

To perform the Correlation Analysis using this filtered data set click on the link Ask another question (above the table)

A selection of 3 tools will appear in the form of questions



Select the second option to explore 'feature correlation based on similarity of expression profiles'



This time the statistically based question(s) being asked is:

'Group my (selected) features according to how similar their expression profiles are'

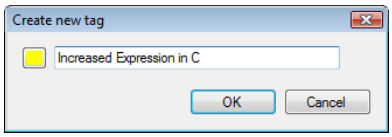
The question is answered by:

'Using Correlation analysis to evaluate the relationships between the (selected) features' expression profiles'.

The answer is displayed graphically in the form of an interactive dendrogram where the distance, horizontally and vertically, between each feature can be taken as a measure of how similar the expression profiles of each feature are.

To highlight all the features demonstrating increased expression in the **'C'** group click on a 'node' for a branch of the Dendrogram (as shown above). As before create a Tag for these features.

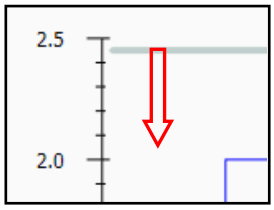
Correlation Analysis enables the grouping of features together according to how similar their expression profiles are.



Clicking on a 'node' in the dendrogram, (as shown on the previous page), causes the features under that node to become selected in the table. Multiple nodes can be selected by holding down **Ctrl** and clicking on the required nodes in the dendrogram as shown below.

Additional features can be selected by holding down **Ctrl** and then clicking on the feature in the table.

Multiple groups can be displayed by setting a 'pruning threshold' on the Dendrogram. To adjust this threshold click and drag horizontal grey the line down over the dendrogram.

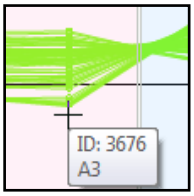


Note: as it moves over each branch, all the branches beneath become colour coded and selected

When you release the mouse button, the expression profiles for each branch of the dendrogram will be displayed. The number and composition of these 'branches' is controlled by sliding the threshold up and down.

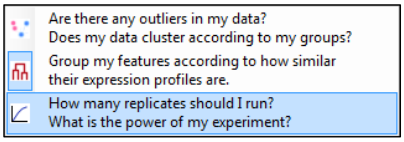
To identify the shape of individual profiles in the expression graphs drag the mouse over the profiles. The current profile is highlighted with a tool tip showing the feature ID and relevant sample information.

Finally, the table can be ordered based on cluster by clicking on the 'Cluster' column.



Power Analysis

To explore the third Statistical analysis of the data click on the blue link Ask another question at the top of the table. The selection of 3 tools will appear in the form of questions.



Select the third option to explore the number of replicates required and obtain a measure of the 'power' of the current experiment.

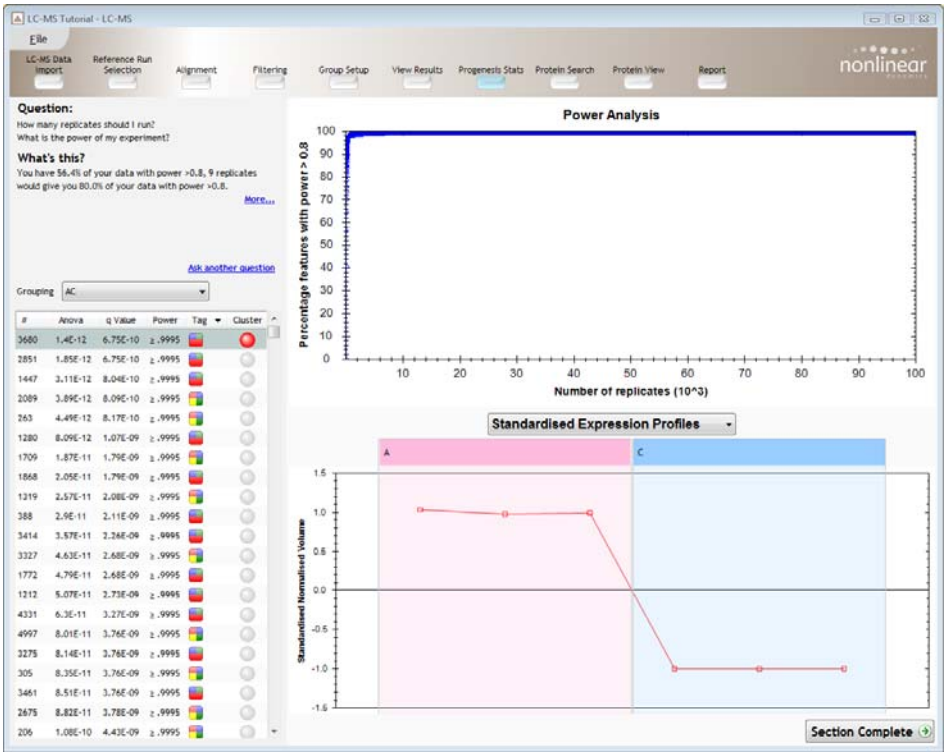
This time the statistically based question(s) being asked of the application is:

'How many replicates should I run and what is the power of my experiment?'

It answers this question by informing you:

'How many replicates you need so that at least 80% of your features with a power >0.8'

Using all **4000 most abundant features** view the power analysis.



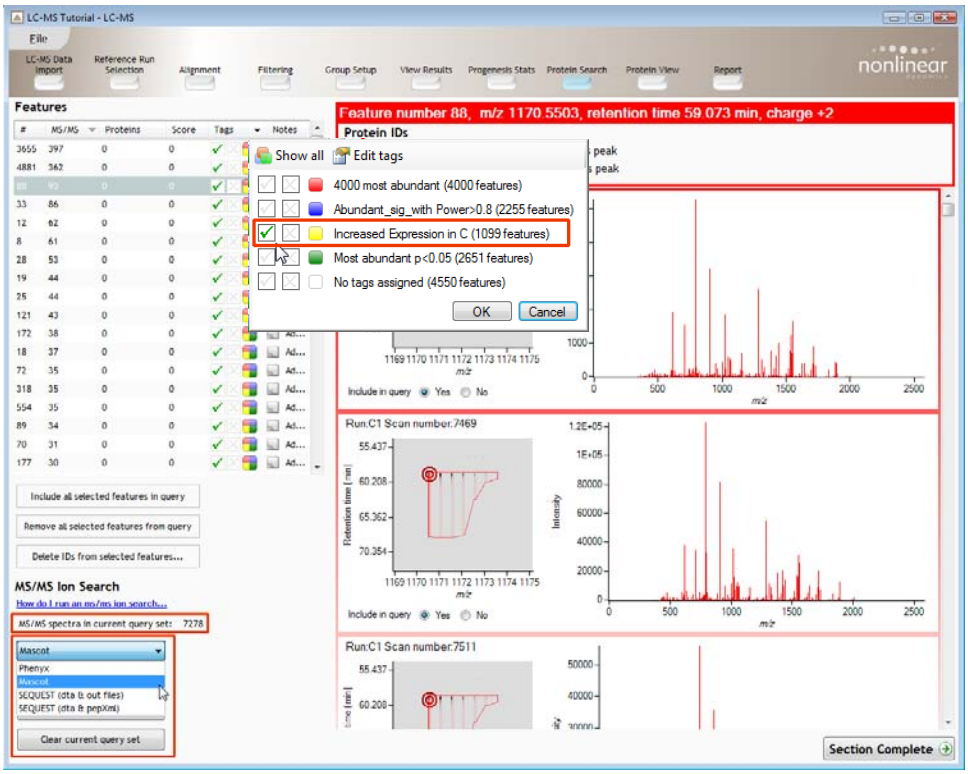
This is displayed graphically showing that 54.6% of the 4000 features have a power of 80% or more and therefore 9 replicates would be required to give you 80% of your data with power > 0.8.

- **The power of a statistical test** reflects our confidence in the experimental data's ability to find the differences that do actually exist
- The power is expressed as a percentage, where 80% power is an accepted level, therefore allowing you to assess the number of sample replicates that would be required to achieve a power of 80%.

To move to the next stage in the workflow click **Section Complete.**

Stage 9: Protein Search

Determining protein ID is dependant on the availability of MS/MS data for the LC-MS runs. This data may be available but limited if the LC-MS was performed in a data dependant MS/MS detection mode due to under sampling. Under these conditions MS/MS data acquisition is dependant on thresholds and parameters set prior to the LC-MS run.



For this tutorial we are using LC-MS runs containing MS/MS data where the data was acquired in a data dependant mode.

Note: by default the table is ordered on the number of MS/MS spectra available for each feature.

Now ensure that the table is only displaying those features showing the behaviour **Increased Expression in C** by setting the **Tags** as shown above, (approx 1099 features are displayed).

Performing an MS/MS Ion Search

Using LC-MS data containing MS/MS spectra, protein ID can be determined for all or a selection of features which have available MS/MS spectra.

1. Highlight all the features with MS/MS spectra in the Features table
2. Click on 'Include all selected features in query'
3. In this case the current 'query set' contains **7278** MS/MS spectra
4. Select appropriate search engine i.e. Mascot
5. Click 'Export current query set' to save search as file
6. Perform search on appropriate search engine and save results file
7. Click 'Import search results', locate results file and open

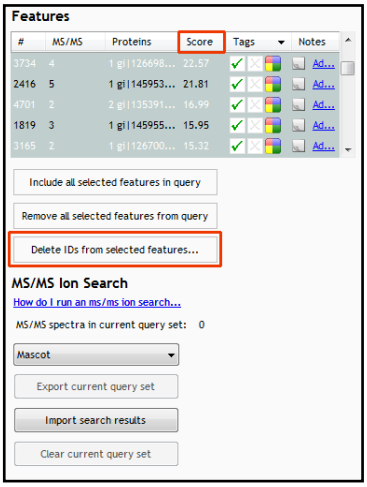
Please refer to Appendix 4 & 5 (page 53) for details of the 'Search Engine' parameters

Note: the blue link tells you the appropriate formats for exporting ID results

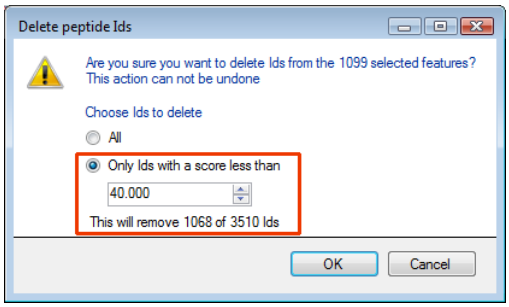
Note: an example Search Results file is available in the folder you restored the Archive to (Protein search results_v2.xml). Select the 'Mascot' method and import this file to see results like those below.

On importing the Search results the Features table updates to reflect the identified proteins and the relevant score for each searched feature.

Now select all the features in the table by clicking on a feature and pressing **Ctrl A**. Order them based on descending 'Score'. Decide on an appropriate threshold, 40 for example.



Click on '**Delete IDS from selected features'** this opens the dialog shown below.



Set the threshold to 40, this removes 1068 ids from a total of 3510

Click **section complete** to move to the Protein View stage.

Stage 10: Protein View

Having filtered at the Protein search stage you can now further validate the protein ids by reviewing the peptides that have contributed to each Protein Identification. This stage allows you to order on the basis of protein Score and then examine the behaviour of the assigned peptides and resolve any conflicts for the various peptide assignments.

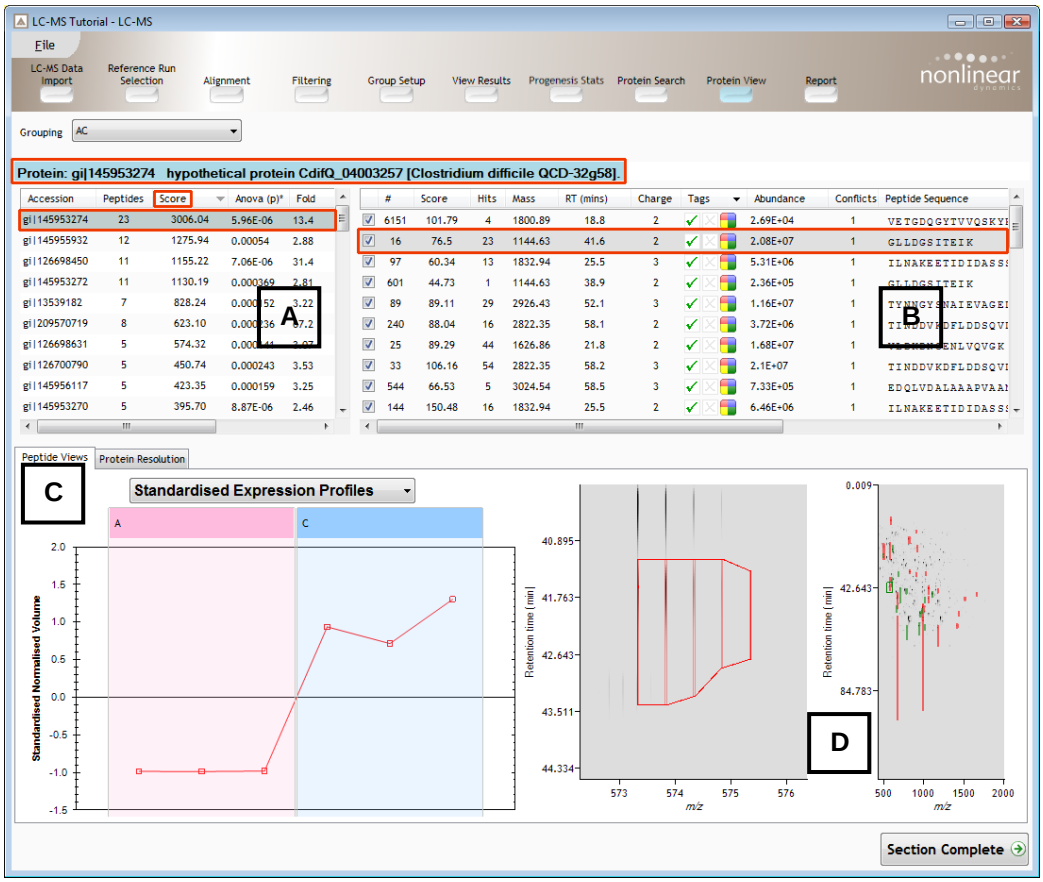
The Protein view provides a number of interrelated graphical and tabular views to assist you in the validation of the peptides that have been assigned to proteins and also to review the relevance of the data returned from the search.

In this tutorial example the organism under study is *Clostridium difficile*

In this tutorial example apply an **Acceptance Criteria**, based on your experimental details to define levels of stringency with regards to database integrity :

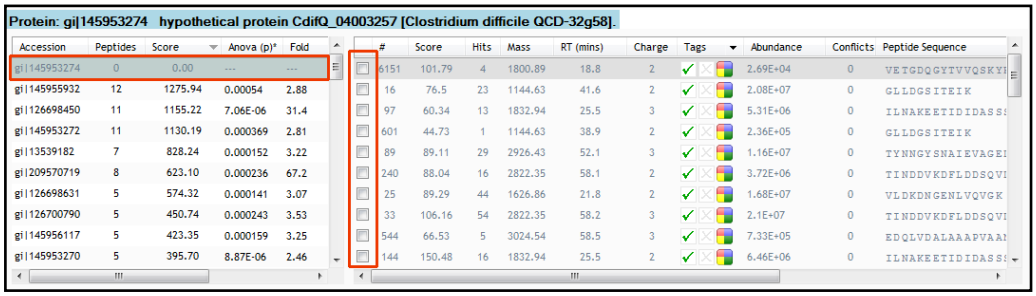
- Resolve conflicts of peptide assignment
- Remove proteins based on relevance threshold (i.e. 'Hypothetical')
- Remove proteins based on biological relevance (i.e. Incorrect organism)
- Remove peptides where only **one hit** is returned

When you open the Protein View order the data in the protein table (A) on the basis of **Score**.



The first protein on the list has 23 peptides assigned, however it is identified in the database as 'Hypothetical', therefore based on our 'acceptance criteria' we want to remove this assignment for all the peptides

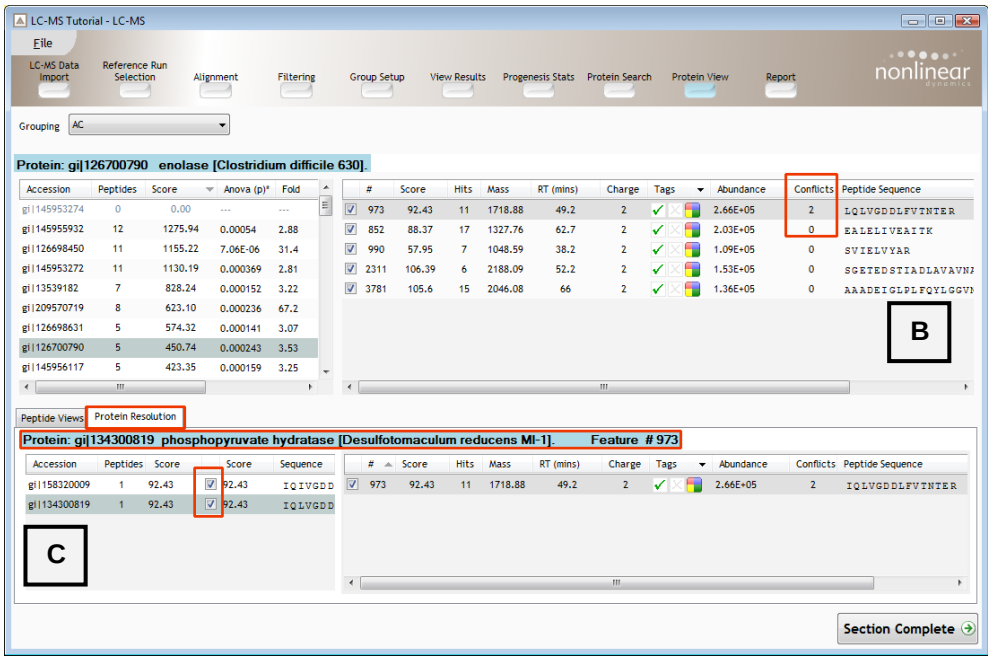
To remove all the peptide assignments for this protein, in the Peptide Table (B) highlight all the peptides by clicking on the first one and then pressing Ctrl A. Click on the tick box, notice how the protein becomes 'greyed out' and the peptide count for this protein is reduced to 0.



You can repeat this for all 'hypothetical' proteins and other inappropriate assignments.

The next example demonstrates how an peptide assignment conflict can be resolved between two proteins.

Select **'enolase'** from the protein table this has a score of 450.74 and 5 assigned peptides. In the peptide table order on the **Conflicts** column, click on the first peptide with 2 conflicts and then on the **Protein Resolution** tab on the Peptides Views (C).

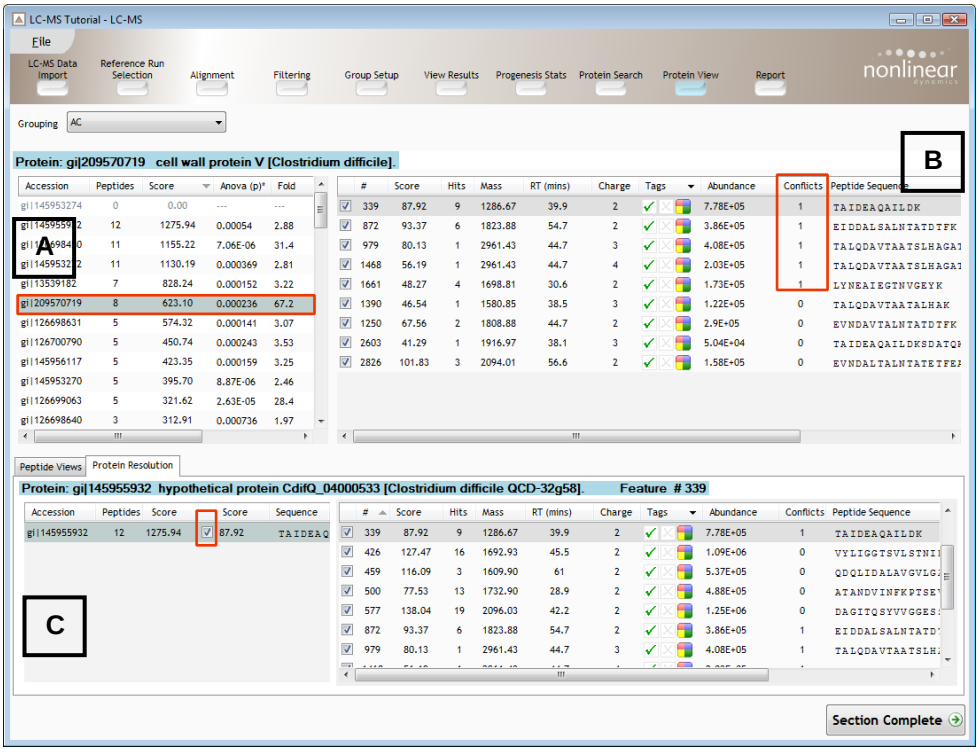


This view displays the 2 conflicting peptide assignments. Not only are the conflicting assignments from the wrong organism but they also contains fewer peptides than the **enolase** protein from *Clostridium difficile*.

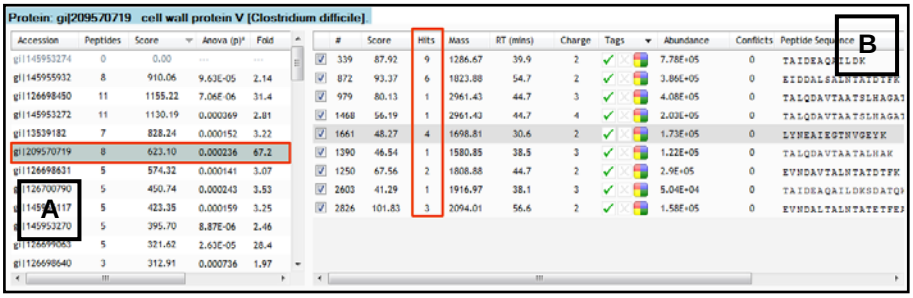
When you 'un-tick' the proteins in the 'Protein Resolution' table then the conflicts for this peptide are reduced to 0 in the Peptides table, and the conflicts are resolved.

Now click on the **'cell wall protein V'** (2 entries above **enolase** in the table). This has 8 peptides currently assigned and score of 623.1.

Order on **Conflicts** in the Peptide Table (B) then using the Protein Resolution table (C) remove the conflicts as shown above.

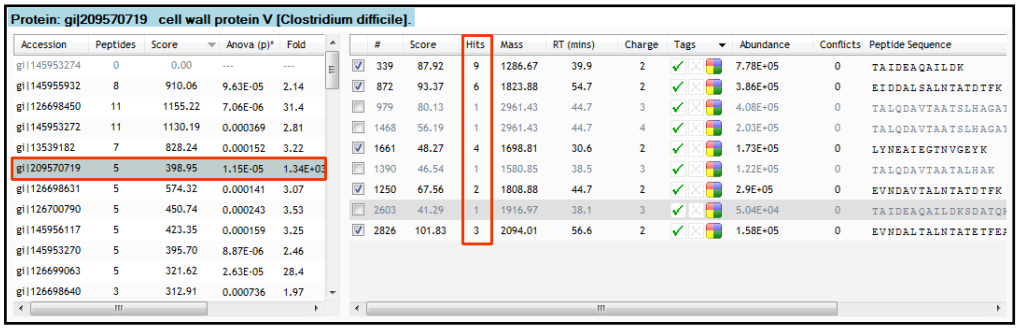


Having removed the conflicts now un-tick all the remaining entries in the Peptide window (B) that have only a single **Hit** for the assigned peptide.



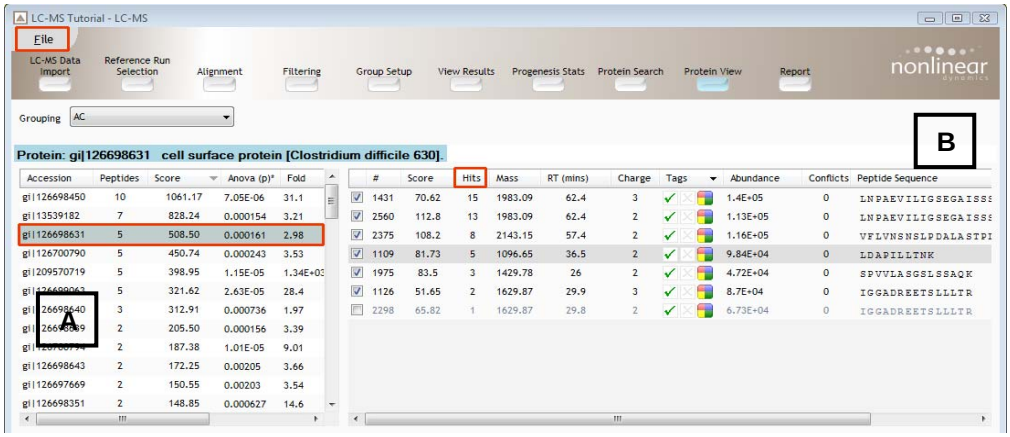
When you un-tick the peptides with only a single **hit** the number of peptides assigned to **cell wall protein V** and the Protein score will update.

Also note the 'Peptides with single hits are now 'greyed out.



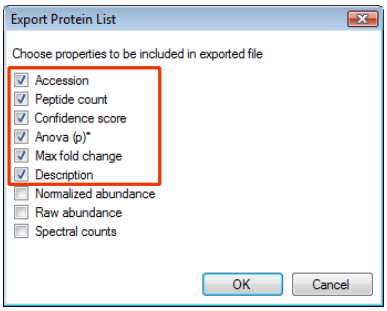
Now as an example of the process use these simple **Acceptance criteria** and the tools provided to rapidly resolve conflicts for all the proteins in the Protein table (A). This should remove all 'Hypothetical' proteins, Proteins coming from any organism other than *Clostridium difficile* and peptide assignments with only one hit.

Finally order the Protein table (A) using descending score, it should look similar to below with the peptides sorted on 'Hits' in the Peptide table (B).

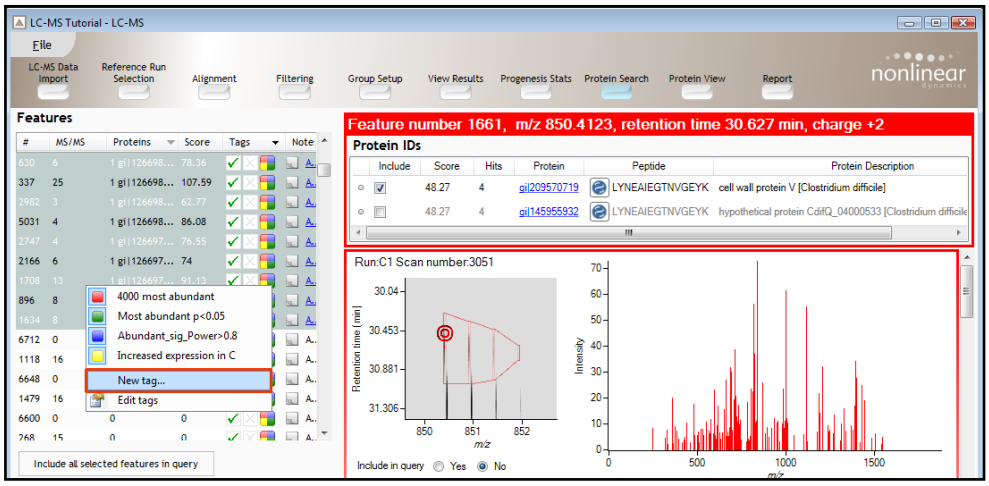


Note: this is a worked example for the demonstration of how the various tools may be applied in the **Protein View**. Hence there are a number of approaches that can be applied to the validation of the data, for example you may wish to switch off proteins that only have one peptide assigned.

Having validated the peptide assignment you can Export the Protein list by selecting this option from the **File** menu. You can control the output (csv format) using the dialog provided

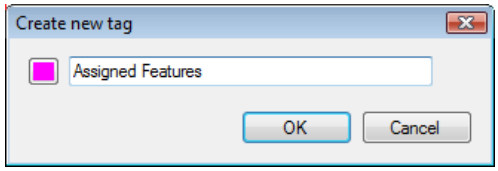


Now return to the **Protein Search** stage by clicking on the icon in the **Workflow** at the top of the screen.



Note: in the Features list, if you have resolved all the conflicts, there will only be one protein assigned to each feature.

Create a new tag for the selected features and call it **Assigned Features**



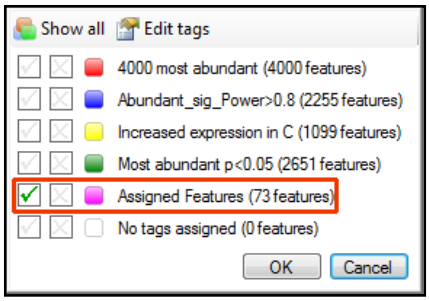
Now move to the Report section by clicking on **Report** icon on the workflow at the top of the screen.

Stage 11: Reporting

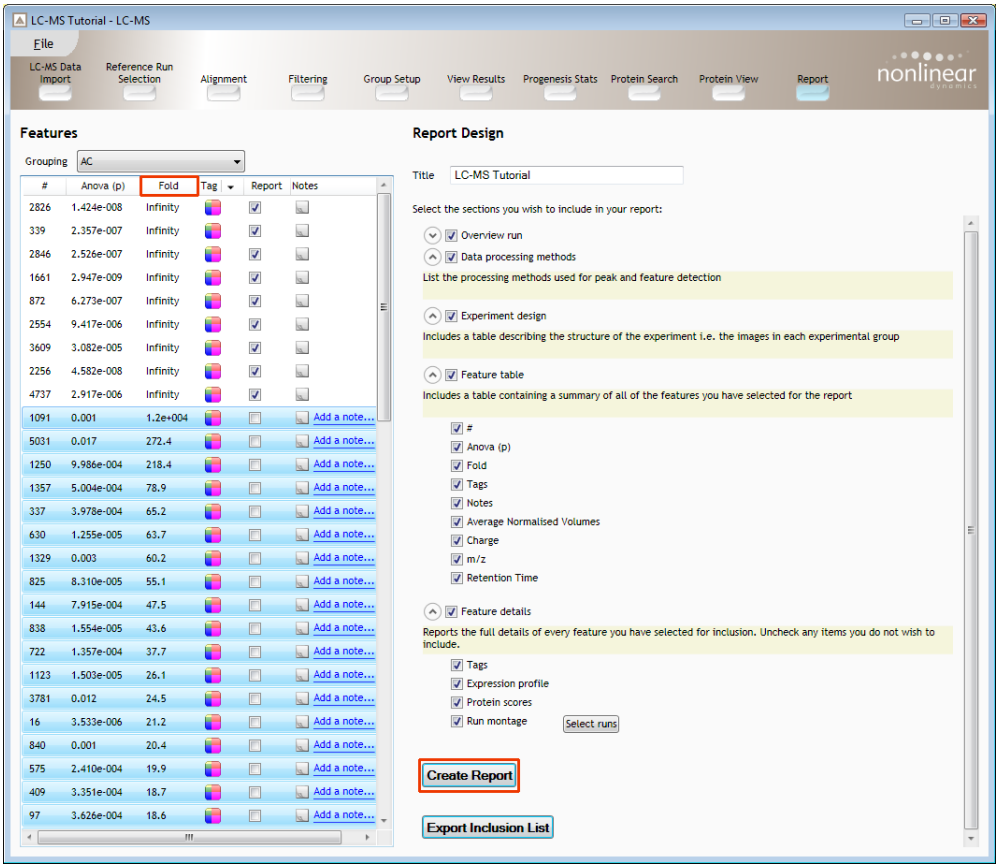
The **Report Design** stage allows you to select what views you want to include in a report based on the list of **currently selected features.**

As an example we will create a report for **only** the features with identified proteins and showing an 'Infinite fold' difference between the groups AC.

1. First reduce the features to report on by selecting the ‘Assigned Features’ tag. In this example it reduces the number of features in the table to 73.

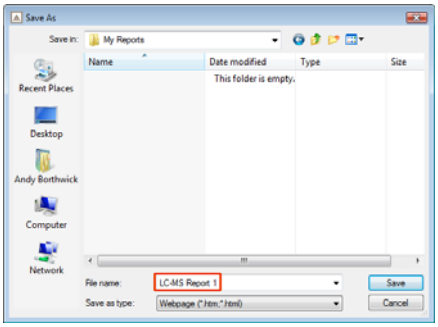


2. Order on Fold then highlight all features with a 'Fold' less than 'Infinity'
3. Click on the tick box to un-select these features
4. Expand the various Report Design options (by default they are all selected)
5. Un-tick as appropriate

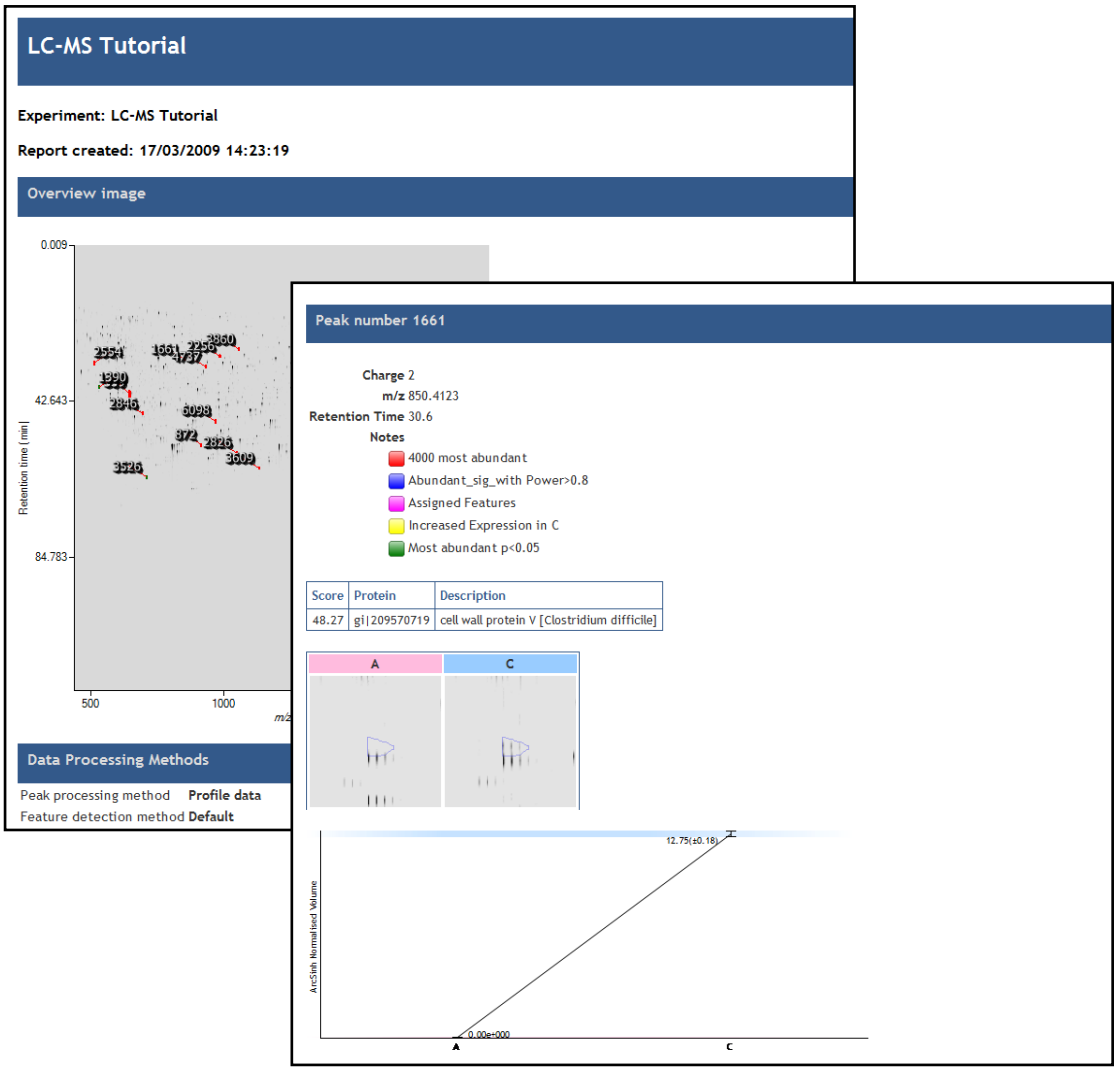


6. Click **Create Report**

This opens a dialog to allow you to save the report, after which it will be opened in the form of a web page.



Scroll down to view details for each feature and montage.

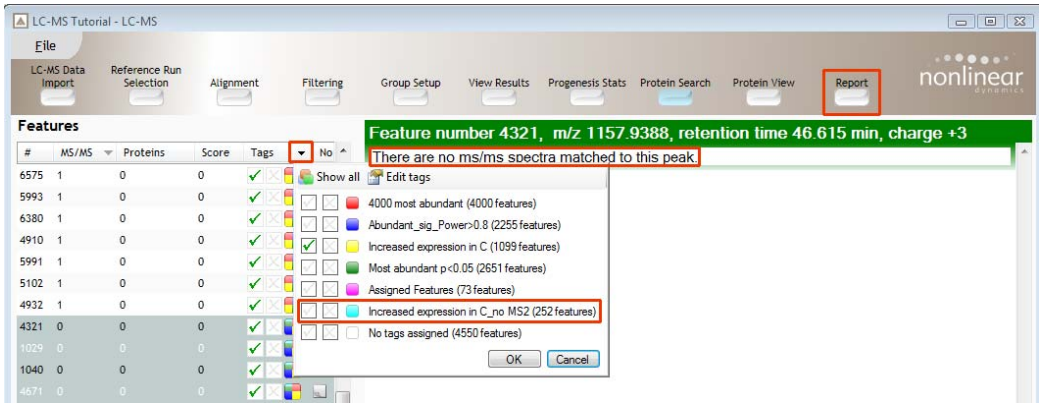


Alternatively the features can be exported as an ‘Inclusion list’.

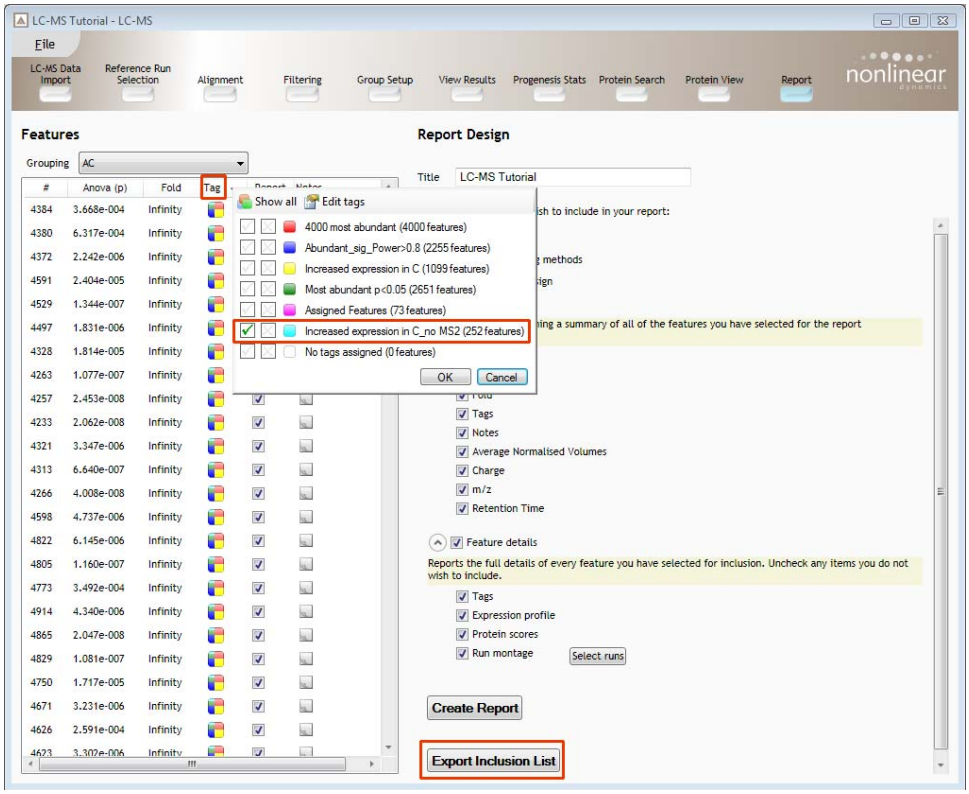
As an example recall those features showing a significant increased expression in strain C with no MS/MS spectra, there m/z and retention times can be used to construct an inclusion list for additional runs 'targeted' at acquiring the missing MS/MS spectra .

By returning to the **Protein Search** stage in the **workflow** you can tag these features (252 in this example). To do this first ensure that only the features shown in the table are those with

'Increased Expression in C' (1099 features) by ticking this tag in the drop down (see below), order the table on MS/MS and highlight those features without spectra. Tag them, **'Increased expression in C_no MS2'**. Now return to the Report stage by clicking on the icon in the workflow.

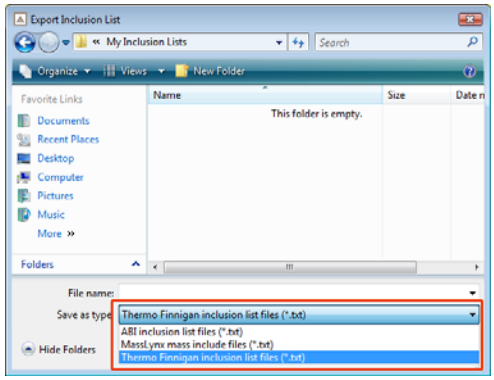


Select the 'Increased expression in C_no MS2' tag from the drop down to display only those 252 features.



At this stage you can further refine the list on the basis of Anova and fold change if required

Finally export the inclusion list in the appropriate MS machine format to use in the acquisition of additional MS/MS spectra from a new LC-MS run.



Note: The new LC-MS runs can then be added to the original experiment to increase the MS/MS coverage using the **Add data files** facility at the Data Import Stage.

Congratulations!

This document has taken you through a complete analysis using Progenesis LC-MS, from Alignment through Analysis to generating lists of interesting features using powerful Multivariate Statistical analysis of the data.

Hopefully our example has shown you how this unique technology, available within Progenesis LC-MS package, can deliver significant benefits with

- Speed
- Objectivity
- Statistical Power

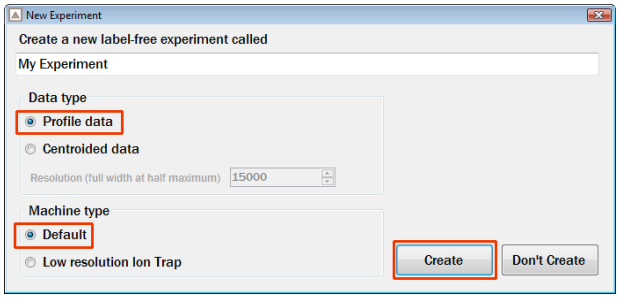
If you would like to see the benefits of running Progenesis LC-MS using your own data and explore the LC-MS Data Import module as well as the rest of the workflow please go to the next section.

Appendix 1: Stage 1 Data import and QC review of LC-MS data set

You can use your own data files, either by directly loading the raw files (Thermo and Waters) or, alternatively convert them to mzXML format first.

To create a new experiment with your files select **New** give your experiment a name. Then select data type, the default is 'Profile data'.

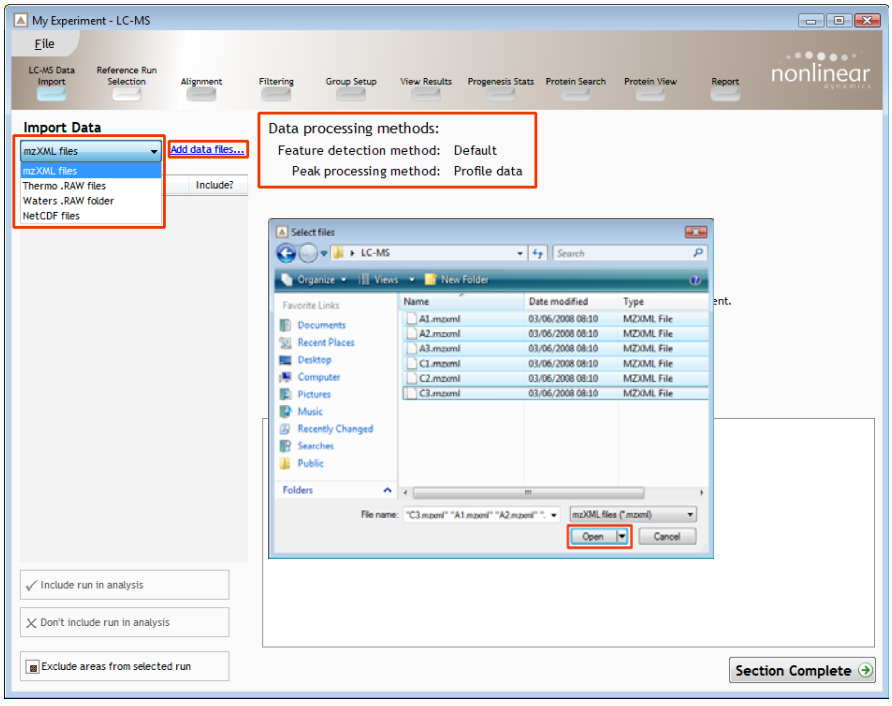
Note: if you have converted or captured the data as centroided then select Centroided data and enter the Resolution for the MS machine used. There is also an option for Low resolution Ion Trap data.



Click Create to open the LC-MS Data Import stage of the workflow.

Select the 'Import Data file format', in this example they are mzXML files

Then locate your data files using the Add data files... link.

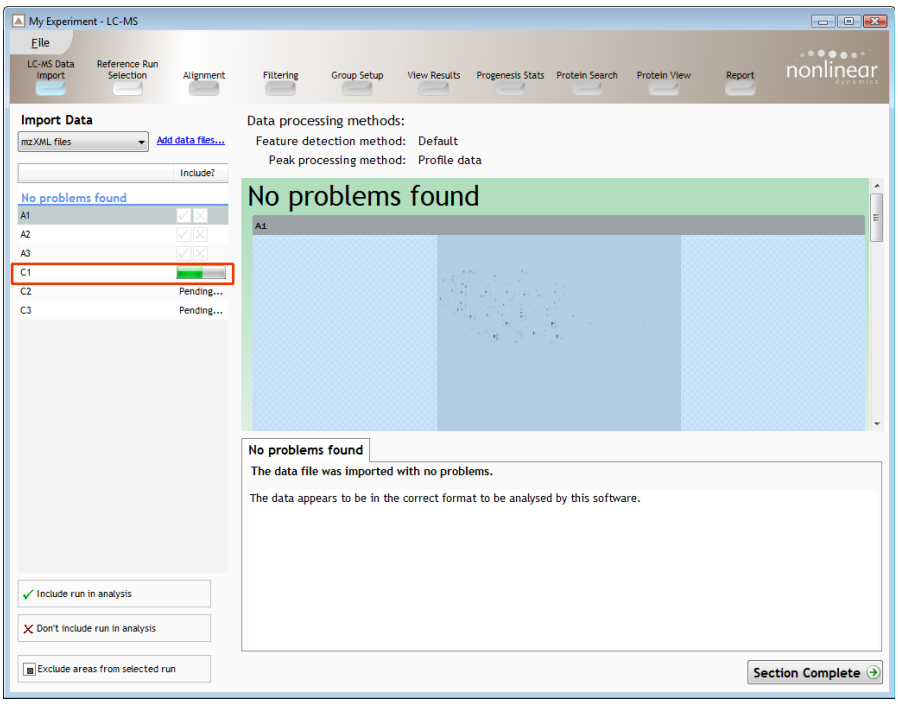


Locate and select all the Data files (A1 to C3).

On loading the selected runs your data set will be automatically examined and the size of each file will be reduced by a 'data reduction routine', which reduces data by several orders of magnitude but still retains all the relevant quantitation and positional information.

Note: For a large number of files this may take some time.

Each data file appears as a 2D representation of the run. At this stage you will be warned if any of the data files have been 'centroided' during the data acquisition and conversion process.

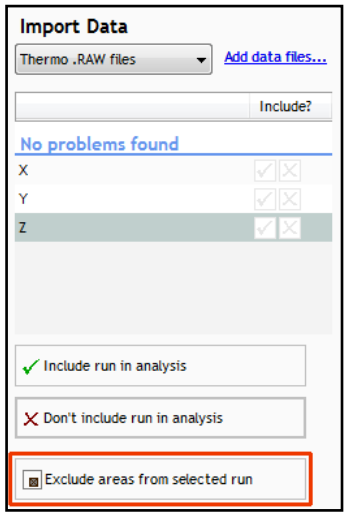


Note: as each data file is loaded the progress is reported in the **Import Data** list. The dialog below the image reports on the QC of the imported Data files. In this case 'No problems found' with the this data file

Now move to the next stage in the workflow (page 6 in this tutorial) by clicking **Section Complete.**

Appendix 2: Stage 1 Data QC review and addition of exclusion areas

During the process of Data QC you may identify areas of the raw data for a particular run that appear 'noisy' yet still have identifyable 'isotopic patterns'.

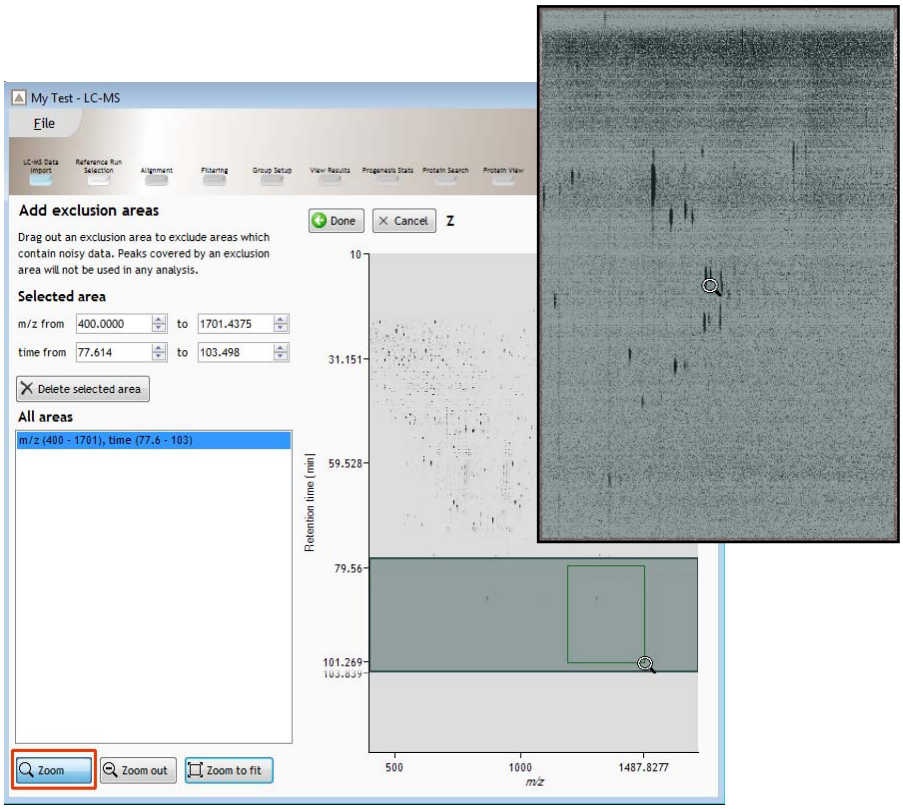


For example if the run is part of a 'replicate set' of runs it is possible to exclude such areas on the noisy run by applying a mask to the area. By doing so this area is excluded during the initial part of the detection process in order that it does not 'interfere' with the detection of the features in the replicate group.

To do this select Exclude areas from selected run on the bottom left of the Software.

Drag out an area over the noisy part of the image to create the mask.

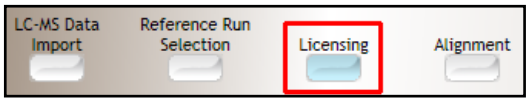
Note: if you now zoom into the masked area using the **Zoom** tool you will see the isotopic features in the noise.



Note: if the level of noise is high and affecting many of your runs a preferred approach would be to re-optimise the chromatography to improve the levels of noise in your data

Appendix 3: Licensing Runs (Stage 3)

When setting up a **New experiment** if you are evaluating LC-MS with unlicensed Runs then the licensing page will open after **Reference Run Selection**



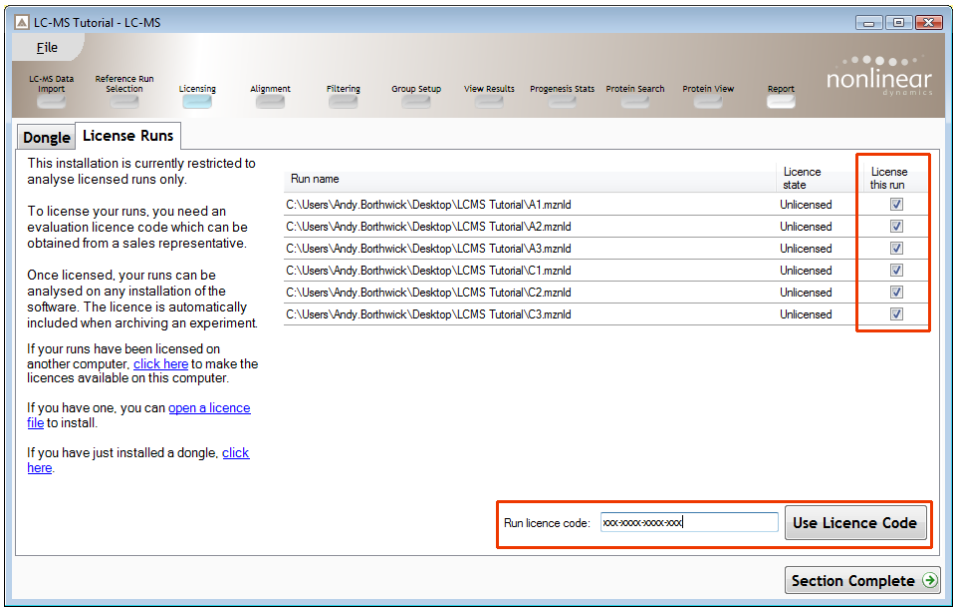
If you already have a programmed dongle attached to your machine then the following License Images page will not appear.

To use this page to License your Runs **you must first either obtain an 'Evaluation' License Code from a Nonlinear Sales Person or have purchased a license code directly from Nonlinear**.

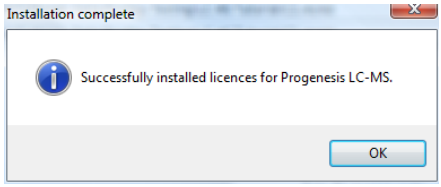
Each code will allow you to license a set number of Runs.

The Runs you wish to License will be listed as shown below.

To activate license(s) for the selected images enter the code in the space provided and click **Use Licence code.**



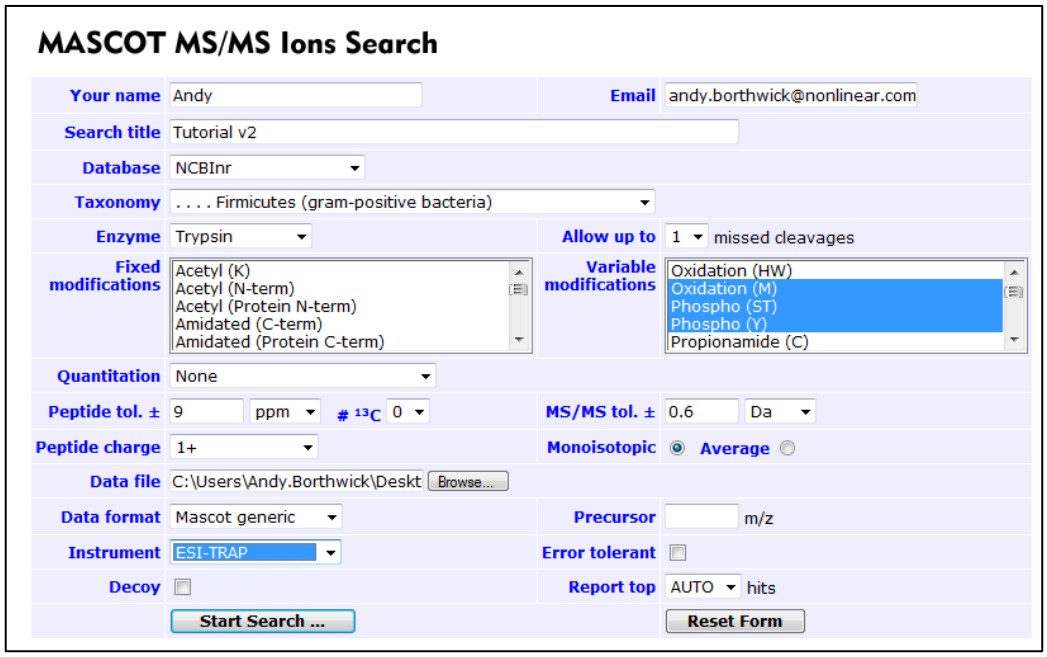
A message confirming successful installation of your image licenses will appear.



Click **OK**, the view will update the and Alignment the next stage in the workflow will open with the licensed files.

Appendix 4: Search engine parameters (Stage 9) Mascot

The parameters applied to the Mascot search that yielded the search results used in this example tutorial are shown below:



Database : NCBInr (circa 03/09) was used with the Taxonomy restriction set to Fermicutes

Variable modifications: Carbamylation(C), OxidationM, Phospho (ST) and Phospho (Y)

Peptide Tol: 9ppm

Instrument: ESI-Trap

Appendix 5: Search engine parameters (Stage 9) Phenyx

The parameters applied to the Phenyx search that yielded the search results used in this example tutorial are shown below:

IDs	60629
Title	
File(s)	C:\Users\Andy.Borthwick\Desktop\LCMS Tutorial\Abundant C.mgf (mgf 108913 Kb)
Databank(s)	NCBInr (20080114)
AC	
Taxonomy	Firmicutes
Scoring Model	ESI-LTQ-Orbitrap (CID_LTQ_scan_LTQ)
Parent Charge	1,2,3,4 (trust=medium)
Round #	1
Modifications	Oxidation_M[variable, <=4] PHOS[variable, <=4] Cys_CM[variable, <=4]
Enzyme	Trypsin_(KR_noP) miss. cleav. 1 cleav. mode. normal
Parent tol.	0.01Da
Pept thresholds	length>=6 score>=6.0 p-value<=1.0E-6
AC Score	6.0
Conflict resolution	yes
Turbo scoring	tolerance=0.5Da coverage >=0.2 series=b;b++;y;y++

Database : NCBInr (circa 03/09) was used with the Taxonomy restriction set to Fermicutes

Variable modifications: Carbamylation(C), OxidationM, Phospho

Peptide Tol: 0.01Da

Instrument: ESI-Trap